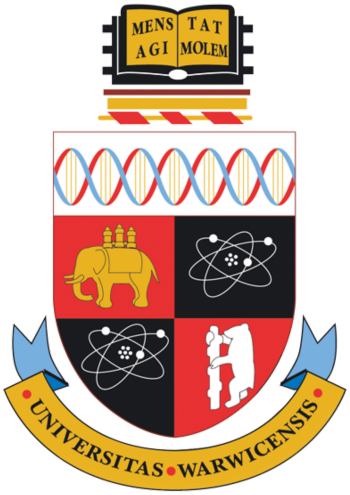
From Words to Mind:

What Can We Learn about Us from the Language We Produced?



by

Ying Li

A thesis submitted in partial fulfilment of the requirements for the degree of

Doctor of Philosophy in Psychology

Department of Psychology

University of Warwick

March 2019

Table of Contents

List of Figures

List of Tables

Acknowledgement

My heart is filled with gratitude when writing this page because throughout the PhD journey I have been accompanied by great mentors, friends and families.

My primary advisor, Thomas Hills, has always been enormously supportive to my personal and academic growth. He sets an ideal exemplar of an intelligent, open-minded, creative, unconventional, modest, and dedicated scholar. He offered tremendous help to my studies and meanwhile left me with great freedom to explore my own passions and interests. I have enjoyed my PhD in a way that that far exceeds my most optimistic expectation, and it is inseparable from Thomas' role as my advisor.

I would like to acknowledge my gratitude to Ralph Hertwig, certainly an unofficial advisor during my PhD and official advisor in the next 3 years, for his wise advises on research and career.

Tomas Engelthaler and Cynthia Siew have been dear friends and amazing colleagues. Together we had a lot fun time hanging out, working on interesting projects, supporting and learning from each other.

My studies were founded by the Leverhulme Trust. I appreciate the institution for seeing the potential in me and giving me the opportunity to turn my interests and ideas into solid work.

The Department of Psychology at Warwick with all its staff has made this institution a home to me. They will forever hold a special place in my heart.

Finally, those to whom this thesis is dedicated: my dear parents and grandparents – 侯闻洁，李增民，张渌水，侯复华，申屠冬花，李运. Their unconditional love, has certainly been the most powerful support along my life journey.

Declaration

This thesis is submitted to the University of Warwick in support of my application for the degree of Doctor of Philosophy. It has been composed by myself and has not been submitted in any previous application for any degree. The work presented (including data generated and data analysis) was carried out by the author except in the cases outlined below.

Inclusion of Published Works

Parts of this thesis have been published or prepared for publication by the author.

Chapter 2 includes

Li, Y., Masitah, A., & Hills, T. T. (In preparation). The Emotional Recall Task: Juxtaposing Recall and Recognition-Based Affect Scales.

Y.L and A.M are both first authors on this paper. Y.L., A. M, and T.H. designed the research; Y.L. and A. M collected the data; Y.L. and A. M analysed the data; Y.L., A. M, and T.H. wrote the paper.

Chapter 3 includes

Li, Y., & Hills, T. T. (In preparation). Quantifying historical change in patterns of immigrant sentiment.

Y.L. designed the research, performed the research, analysed the data; Y.L. and T.H. wrote the paper.

Chapter 4 includes

Li, Y., Hills, T. T., & Hertwig, R. (In preparation). A cultural history of risk.

Y.L., T.H., and R.H designed the research; Y.L. performed research; Y.L. analysed the data; Y.L., T.H., and R.H. wrote the paper.

Chapter 5 includes

Li, Y., Engelthaler, T., Siew, C. S., & Hills, T. T. (2019). The Macroscope: A tool for examining the historical structure of language. *Behavior Research Methods,* 1-14.

Y.L. designed the research; Y.L. performed the research; Y.L. analysed the data; Y.L., T.H., C.S, and T.E wrote the paper; Y.L and T.E built the website.

Abstract

Language is not just a record of past events. It represents our interactions with the environment: how we feel, conceptualise, construct and communicate our experience. It also weaves the cultures within which our identities, emotions, values and a long list of other important psychological phenomena are shaped. This makes language a fertile ground for studying psychology. This dissertation shows how quantitative text analysis informs emotions of individuals, opinions in a society, and a history of a concept. I suggest that text analysis, a thread of research that dates all the way back to the earliest days of psychology, should revive in light of the availability of many unprecedentedly large corpora and extend its scope beyond case studies of individual minds. The Macroscope, a linguistic tool we developed for examining the historical language structure, makes it convenient for anyone to explore and investigate historical change of psychology in the context of socioeconomic dynamics.

Chapter 1 : From Words to Mind

1.1 Prologue

Human beings have been living in a dual reality: on one hand, the objective reality of land, rivers, and sky, shared with other living creatures, and on the other hand, the imaged reality of nations, religions, commercials, etc. The imagined reality exists in shared understanding of concepts. For example, *law* has power because members of community believe, understand, and act upon *law* and its related concepts within a judicial system. This imaged reality has constituted a major part of human society and it is only possible with command of complex language (Harari, 2014). It is hardly an exaggeration to claim that language reflects, influences, or even constructs the reality we live in (Dunbar, 1996). This immediately makes language a useful resource to study the psychology and action of its users at individual, collective, or historical levels. This thesis, as part of the historical thread of text analysis, shows how meanings can be extracted from language to shed light on individual mental states, public opinions, and cultural change.

How language represents meaning is a question that demands a clear answer before any attempt to extract meaning from language can be undertaken. Many theories have been proposed but unfortunately little consensus has been reached. In this section, I aim to review important theories of meaning and organise them to show how one theory disagrees with, develops from, or complements another. I will then discuss how these theories relates to the methodologies used in this thesis. Before diving into theoretical details, basic concepts important to language studies need be explained.

1.2 Linguistic sign and its properties

Language can have meaning in two fundamental ways: through what is referred to as an encoded sign (semantics), and through what it does in context (pragmatics). Every meaning production uses two elements: a linguistic sign (sound or word) and its referent (concept). A sign only has meaning when members of the speech community agree to that meaning (Kramsch, 2009). Linguistic signs are mostly arbitrary, amodal, and abstract (Hockett, 1960, Glenberg & Kaschak, 2002; Pecher & Zwaan, 2005; Zwaan, 2014). They are mostly arbitrary because there is no clear fixed one-to-one mapping between signs and their references. That means, few information in the form of a sign (such as morphological features) informs us about

their meaning. Signs are also mostly amodal and abstract. They are amodal because they are not related to any sensory perceptions; and they are abstract because they do not refer to any specific object (even words as concrete as *bulldog* refer to a concept/category that encompass many individuals). We acknowledge that a few words (such as onomatopoetic words such as *tweet*, *click* and *bang*) are to certain extent iconic, meaning that a word form bears some resemblance to its meaning. Although such iconicity offers advantage in language processing and learning (Dingemanse et al, 2015), most words are arbitrary and amodal, probably because iconic words are too specific to contexts and referents and therefore a more iconic language would make it more difficult to express and learn abstract concepts.

Table 1.1 Theories of meaning

	Theories	**Key features**
1	Feature-based theory (Katz & Fodor, 1963)	A checklist-based approach.
2	Prototype theory (Rosche, 1973)	Emphasis on subjective perception. Categories exhibit family resemblance.
3	Frame semantics (Fillmore, 1975, 1976)	Words represent categorisation of experience
4	Distributional semantics (Firth, 1957 Burgess & Lund, 1997; Landauer & Dumais, 1997)	Statistical distributions of linguistic forms per se represent knowledge. "One shall know the meaning of a word by the company it keeps" (Firth, 1957).
5	Language and situated simulation (Basalou, 2008)	Two systems, one perceptual and one linguistic functions together to represent meaning.
6	Deacon's hierarchy of signs (Deacon, 1997)	A hierarchical structure of iconic, indexical, and symbolic processes explains language processing.
7	Symbol Interdependent Theory Louwerse (2008)	Language encodes perceptual information.
8	Osgood's psychological meaning of words (Osgood, Suci, & Tannenbaum, 1957)	Not a formal theory of meaning Compare thousands of words along the same quantifiable scale of psychological features.

1.3 Theories of meaning

An intuitive question is why not simply learn meaning from a dictionary. Dictionaries give words meaning using other words. In this sense, since dictionaries provide paraphrases rather than meaning, what they offer is essentially relations between linguistic signs. One must reply on prior knowledge of some words to understand others. Without any prior knowledge,

as Searle (1980) illustrated in his Chinese Room thought experiment, it is not possible to learn meaning of language simply through relations between signs. Theories of meaning must explain how mind attaches meaning to signs rather than a description of what signs refer to. In the theories introduced below, the word *meaning* is often used interchangeably with its closely-related concepts such as *concept* (fundamental basic unit of knowledge), *category* (concept with members) or *knowledge*.

Feature-based theory. During the 20th century, discussion on theories of meaning has shifted from a traditional definition approach that relies on necessary and sufficient conditions as in Aristotelian logic to a more subjective, context-based approach with heavy emphasis on contiguity between meaning and experience (Barsalou et al, 2008). An exemplar of the definition approach to meaning is Feature-based Theory (Katz & Fodor, 1963) that attempts to explain concepts in terms of the process of understanding how they are organised into categories. It defines a category by a checklist of essential indispensable attributes. The checklist approach is motivated by treating "categories as logical bounded entities, membership in which is defined by an item's possession of a simple set of criterial features, in which all instances possessing the criterial attributes have a full and equal degree of membership" (Rosch & Mervis, 1975).

A common critique to the Feature-based theory is that categorical membership is not an all-or-none phenomenon. Concepts, and the reality they refer to, often do not have clean-cut boundaries (Mervis & Rosch, 1981). For example, it is difficult to demonstrate a clear borderline between different colours: no single line can be drawn in the spectrum to separate where red stops and orange begins.

Prototype theory. More recent views on concepts such as Prototype Theory (Rosch 1973, 1975, and Mervis & Rosch 1981) embrace the idea of *family resemblance* (Wittgenstein, 2009), which states that members of a concept are connected by a series of overlapping similarities, where no one feature is common to all members. Wittgenstein (2009) used *game* as an example to demonstrate that words have no definitive meaning since no single thing is common to all activities represented by the word *game*. As a radical departure from the traditional definition-based model of concepts, Prototype Theory argues that categorisations are made based on perceived similarity to a prototypical model of the category, which is formed by aggregating all the objects in the category one has previously encountered. This suggests categories cannot be defined by a single set of criterial attributes. Members of a category are semantically structured in a form of radial network with the prototype in the centre. Consequently, they are not equally representative of the category. For example, a robin is

usually perceived to be more prototypical of a bird than an ostrich or a penguin (Malt & Smith, 1984).

Frame semantics. Prototype theory involves previous experience in forming a prototypical model. Fillmore (1975, 1976), with greater emphasis on contiguities between a word and how it has been experienced in its underlying cultural context, proposed Frame Semantics. He argued that we think, largely unconsciously, in terms of conceptual frames – mental structures that organise our thought. He further argued that the meaning of every word is mentally defined by elements of organised mental structure of experiences, which he called "frames". He demonstrated this idea by studying semantic fields, groups of related words such as *buy*, *sell*, *goods*, *price*, *cost.* The meaning of any one of these words, say *sell*, depends on understanding the frame of "commercial transfer", which, apart from the act of selling, comprises act of buying, seller, buyer, money, transaction, goods, and so on. These are named the basic "semantic roles" – the conceptual elements of the frame.

Fillmore (1975, 1976) argued word meaning can be explained by clarifying reasons a speech community has for creating and using the category represented by the word. Therefore, study of meaning became the study of (1) which frames we use in categorising our experience; (2) what scenarios and semantic roles define each frame; (3) how frames relate to one another. Semantic Frames has provided a powerful device to understand a wide variety of linguistic phenomena. For example, conceptual metaphors can be viewed as frame-to-frame mapping: ways to understand experience constructed under one frame in terms of another (Lakoff & Johnson, 1980). Polysemy arises from alternative frames of the same lexical unit. Same situation can be presented within different framings to serve different purposes (e.g., in discussion on abortion, choice of words *fetus* vs *baby* evokes completely different frames with opposite political implications: frame of medical procedure vs frame of murder). Frame Semantics suggests that the meaning of a word exists not within itself, but in relation to the cultural context in which it is understood, used, and acted upon. In short, words represent categorisations of experience.

Osgood's psychological meaning of words. Different from all previous formal theories of meaning is Osgood's idea of the psychological meaning of words (Osgood, Suci, & Tannenbaum, 1957). He developed a technique for measuring the connotative meaning of words, known as semantic differential. Osgood et al. (1957) shows in experiments that people generally shared similar perception of a list of word properties such as good/bad, hot/cold (Figure 1.2). Corresponding to the emotion theories that propose valence (pleasantness) and

arousal are two major, mutually independent components of emotion experiences (Wundt, 1905; Russel, 1980), Osgood et al (1957) found meaning of words fall along (at least) three dimensions: pleasantness, arousal, and control using a factor analysis of a large number of scales evaluating people's responses to various items. Emotion valence has been found to be a part of a word's lexical representation (Fazio, Sanbonmatsu, Powell & kardes, 1986; Houwer & Randel, 2004; Vigliocco, Meteyard, Andrews, & Kousta, 2009), while formal approaches such as Feature Based Theory usually don't consider emotion valence to be part of a word's lexical representations. Relating to the Osgood's idea is the development of *Affective Norms for English words* (ANEW; Bradley, & Lang, 1999), which contains a collection of 1,034 words rated on valence, arousal and dominance. More recently, Warriner, Kuperman and Brysbaert (2013) extended ANEW to 13,915 words. These databases. Other than sentiment, other psychological properties that have been made into norm dataset are concreteness (Brysbaert, Warriner, and Kuperman, 2014), age of acquisition (Kuperman, Stadthagen-Gonzalez, and Brysbasert 2012), meaningfulness (Paivio, Yuille, & Madigan, 1968), humor (Engelthaler & Hills, 2018), etc.

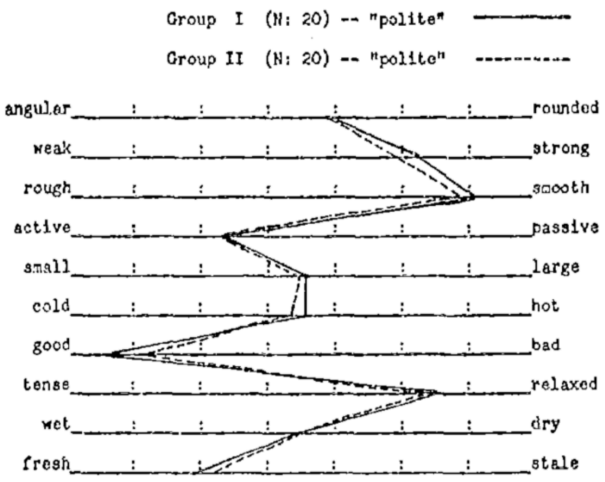


Figure 1.1. Visualisation of the word *polite* in a high dimensional space. The values are means as rated by two independent groups of participants (n = 20). Reprinted from: Osgood, C., Suci, G., & Tannenbaum, P. (1957). The measurement of meaning. Urbana, IL: University of Illinois.

These norm datasets do not attempt to reach a comprehensive meaning of words. Instead, they use words as stimuli to activate one type of perceptual simulations such as image, touch, smell, taste and sentiment, extract the quantifiable part of the simulation, and frame it along a fixed scale (e.g. ask participants to rate on 7-point scale to indicate how angular-round the word is instead of asking for description of an image the word provokes). Those

psychological meanings are not aimed to encode sufficient semantic relationships to produce conceptual mapping like Figure 1.1. Instead it makes thousands of words comparable along the same dimension of connotative meaning. These norm datasets have proven fruitful across many research fields (for a few examples: Dodds et al, 2015; Alhothali & Hoey, 2015; Hills & Adelman, 2015; Hills, Adelman, & Noguchi, 2016).

Embodied vs symbolic approach to meaning. From around the 1970s, the debate between two major camps to understanding meaning started and still lingers around now: embodied cognition, which emphasizes the importance of perceptual, motor, and emotion experiences in our conceptual structure and word meanings and symbolic account which emphasize on the symbolic representations and proposes statistical distributions of linguistic signs per se represent meaning (Burgess and Lund 1997; Landauer and Dumais 1997). Roughly, the debate centres around the question, as Louwerse (2018) summarised, whether "one shall know [the meaning of] a word by the [linguistic] company it keeps" (Firth, 1957) versus "one shall know the meaning of a word by the perceptual simulations it generates."

Symbolic account of word meaning. Major recent breakthroughs in the field of natural language processing have made a strong assumption of the Firthian idea of distributional semantics: words with similar linguistic distributions (used in similar contexts) have similar meanings. Words are represented by vectors of linguistic distributions (or sometimes called word embeddings) that capture semantic and syntactic similarities. For example, word embeddings trained by algorithms such as Word2Vec (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013) and GloVe (Pennington, Socher, & Manning, 2014) are able to capture analogical relationships such as *king is to queen as father is to mother*. Figure 1.1 shows that when projecting high-dimensional word embeddings into a two-dimensional space, a clear categorical mapping emerges: words of similar categorical membership appear to be grouped together. This suggests that when a person sees words in Figure 1.1 presented in a natural language context, by applying statistical language learning alone (Saffran, 2003), they are able to have a sufficiently good idea of how these words are similar to each other.

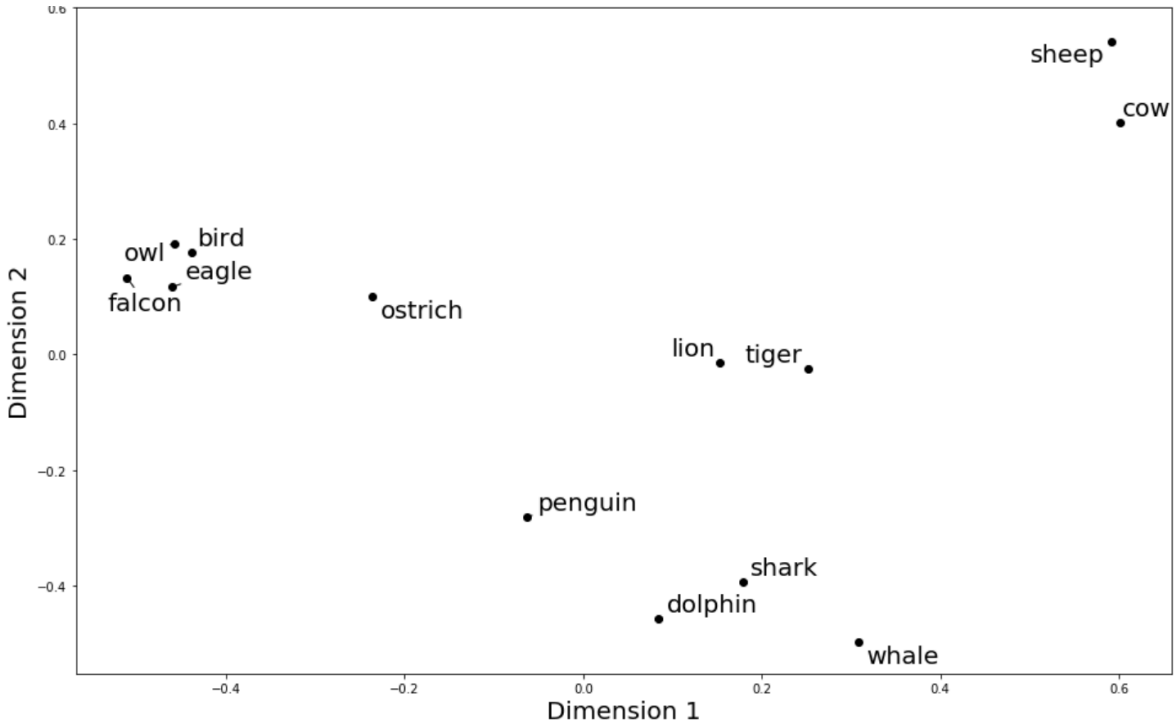


Figure 1.2. Principal Component Analysis (PCA) plot of word vectors of the 300-dimensions (proportion of variance of the scaled data can be accounted for by PCA procedure: R-square = 0.43). The figure is generated using the Macroscope (Li et al, 2019), a linguistic tool that offers analysis on historical language structure (for more details refer to chapter 5).

Embodied account of word meaning. On the other hand, evidence for embodied representation has accumulated showing language processing involves activation of non-linguistic perceptual simulation (Connel & Lynott, 2011, 2016; Louwerse & Connell, 2011; Barsalou, 2008; Pecher & Zwaan, 2005; Semin & Smith, 2008; Shapiro, 2014, for overviews). It has been shown that the brain captures modal states during perception, action, and introspection, and then later simulate these states to represent meaning (Pulvermuller 2002; Barsalou, 2003, 2008; Glenberg, 1997; Damasio, 1989). For example, when perceiving and interacting with dogs, the brain captures modal states in visual, auditory, and somatosensory system on how dogs look, sound, feel, smell, etc. In addition, the brain also captures modal states in actions and introspective states such as affect. Later the brain reactivates these multi-dimensional states, often partially, to represent the meaning of dog.

Unified approach to meaning. Evidence from both embodied and symbolic approaches calls for a unified account. Barsalou (2008) proposed Language and Situated Simulation (LASS) according to which two systems, one linguistic and one perceptual, operate interactively in the process of representing knowledge. On perceiving a word, the linguistic system is immediately activated to recognise the cue word and generate associated conceptual

information. Almost at the same time, both the cue word and its mental associations begin to activate perceptual simulations. Concepts activated by the linguistic system serve as pointers to simulations useful to represent the cue word's meaning. Barsalou argues that perceptual simulation is usually situated to prepare an agent for appropriate actions suited for a specific situation. The degree of involvement of two systems varies across situations. When a superficial linguistic processing strategy is sufficient to perform the task at hand (e.g. in lexical decision and synonym tasks), processing may rely mostly on the linguistic system and little on simulation. On the contrary, in difficult tasks (e.g. verifying that an abstract concept applies to a picture) when linguistic processing is inadequate, the simulation system must be consulted for deeper conceptual information (Wilson-Mendenhall, Simmons, Martin, & Barsalou, 2013).

The two-system view to meaning is also implied in Deacon's (1997) work on hierarchy of signs and Louwerse's (2018) Symbol Interdependence Hypothesis. They suggest that systems of signs are organised in a non-arbitrary manner: the language system encodes perceptual relations (as illustrated in Figure 1.1). Therefore, with a few words' meaning grounded in perceptual experience, meaning then spreads through the network of indexical relations. If one does not know the meaning of *eagle,* by grounding other words such as *bird, falcon, owl*, *penguin* and *ostrich,* the semantic meaning of *eagle* can be bootstrapped through its relationships with other words such as falcon and owl: *eagle* is very likely to be a kind of bird that flies.

It seems both systems are required to represent meanings. With linguistic systems alone, one can only infer relationships among words (*whale* and *shark* are two closely related concepts), without understanding of their actual meaning (what exactly is a *whale/shark*). Concrete words like these must be grounded in perceptual experience (how they look, feel, smell, move, etc) to attain meaning. However, if only perceptual simulations are involved, abstract concept words become extremely difficult to learn because they can't be grounded (Barsalou, 2010). From operations of the linguistic system, abstract words can become meaningful through indexical relationships with other words that can be grounded (Peirce, 1931; Barsalou, 2008; Schwanenflugel, 1991). Language statistics allow for extracting meaning from words using only limited grounding (Louwerse, 2008).

To summarise briefly how the two systems complement each other, the perceptual system grounds meaning of concrete words in perceptual experience, while linguistic and perceptual systems together weave all words into a structured network based on indexical

relationships inferred either from linguistic co-occurrence or from spatial and temporal contiguity. This network functions as a medium for meaning to spread from grounded words to un-grounded ones.

1.4 From word meaning to corpus meaning: less could be more

This review of theories of word meaning suggest richness of possible sources of word meaning and the difficulty of establishing a consensus. However, there are several overarching lessons. First, meaning is subjective, relying on users and its context. Consequently, the meaning of some words can change throughout history and reflect potentially the changing psychology of its users. Second, the semantic and syntactic relationships of words can be extracted from word co-occurrence statistics. Third, the meaning of words is derived from perceptual simulations (Barsalou, 2008), which could include mental images, feelings, touch, smell, taste, etc. Therefore, words that simulate same kinds of perceptions can be compared along the shared dimensions such as valence and arousal (Warriner et al, 2013).

In my thesis, the meaning of a single word is represented by its relationships with other words. This Firthian approach to word meaning is certainly not sufficient to explain how meaning of words is learnt and understood. However, it is not the goal of this thesis to discuss the cognitive mechanism of how the meaning of individual words is processed. Instead, this thesis uses language as a window to understand the psychological states of its producers. Therefore, choosing a Firthian approach incurs little cost from neglecting the grounding problem and meanwhile brings a huge benefit that meaning of a word can be conveniently quantified and comparable to other words. The relations between two words can be represented as either semantic similarity or contextual co-occurrence. Figure 1.3a is an example of the semantic similarity between words. It shows how the historical meaning of *broadcast* (year 1850 and 2000) is related to its corresponding synonyms in semantic space; while Figure 1.3b shows multiple clusters of contextual words with which the word *nuclear* was used. Since language encodes perceptual information and perceptual information can be bootstrapped through linguistic statistical regularity (Louwerse, 2018), a distributional semantic approach can offer a good-enough (though not complete) representation of word meaning (Ferreira, Bailey, & Ferraro, 2002). In addition, a specific dimension of word meaning (e.g. valence) can be extracted to answer specific research questions. For example, when studying how well words produced by individuals describe their general affect (chapter 2), I extracted the valence of each word because this dimension of word meaning is the most relevant to affect.

Intuitively, analysing a single word (e.g. analysis of *risk* in chapter 4) is very different from analysing large corpora (e.g. analysis of immigrant corpora in chapter 3) that consist of thousands of documents created for different reasons, with many authors and across a wide variety of genres. However, they share more similarities than differences. The meaning of a word cannot be studied in isolation. Instead, a comprehensive word meaning can only be discovered from various contexts in which the word has been used. Studying the meaning of a word requires studying a collection of its co-occurring words, which is by definition a corpus. Therefore, analysis used in chapter 3 (study of the meaning of *risk*) and 4 (study of an immigrant corpus) is quite similar: both use a topic model to extract relevant topics, and compare related concepts along certain dimensions of meaning such as valence or concreteness.

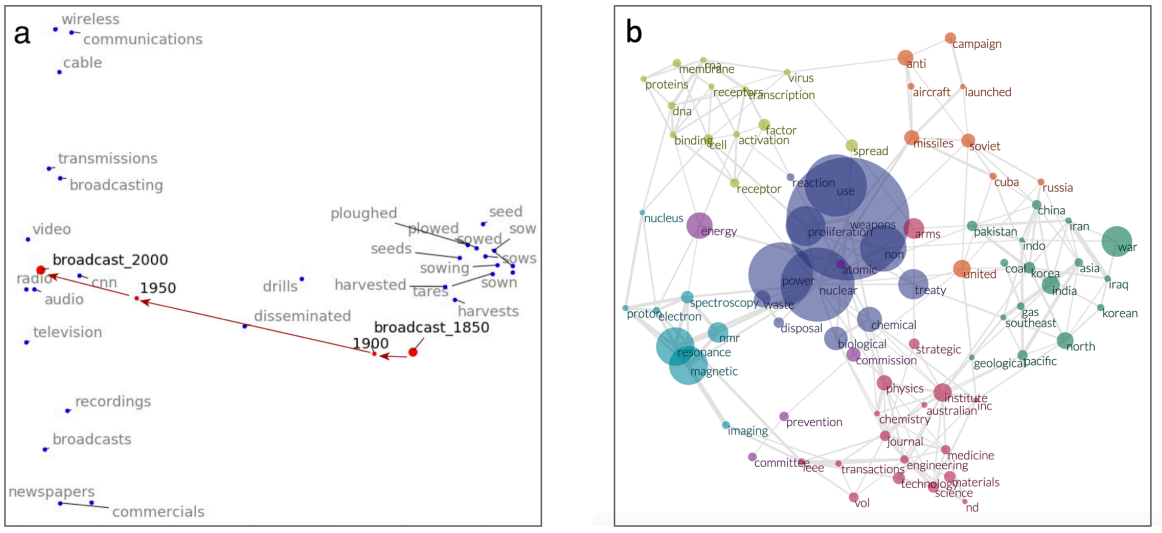


Figure 1.3. Both figures are produced from Macroscope (Li, Engelthaler, Siew & Hills, 2019). (a) Semantic drift of word broadcast from 1850 to 2000. These words are positioned according to their semantic relationships: semantically similar words are close to each other. (b). Contextual network of nuclear in the year 2000. The nodes represent words that appear in the same context as nuclear. Edges are between words if their co-occurrence frequency exceeded a predetermined threshold. The size of nodes is proportional to their usage frequency in a given year. The colors represent the community structure of nodes in the network and each community is represented with a different color.

Corpus meaning can be derived from both a microscope or "macroscope" approach. One can either take a close examination by decomposing the corpus into sub-topics using topic models or contextual network analysis (discussed in greater details in chapter 3-5) or take a bird's eye view by aggregating all words along certain dimension of meanings to retrieve corpus features such as sentiment. Although the latter does not encode information as detailed as the former, it summarises corpus meaning at higher abstract levels and provides quantitative patterns ready for statistical analysis. Given that many attempts have been made to use naturalistic language data to inform psychology constructs, I will next summarise two

principles that seem to guide their methods.

First, since a corpus as large as Google Ngram Books often contains a lot of irrelevant information that contributes more noise than insights to the question of interest, reducing the corpus to a smaller and more relevant subset is essential. Theoretical research questions should guide the development of criteria to decide what information would be singled out for careful analysis. Those criteria are often underlain by the assumptions on how a psychological construct is related to the language usage. For example, Greenfield (2013) and Uz (2014) used pairs of words (such as *choose* vs *obliged, get* vs *give*, *act* vs *feel, I vs we*) to index individualistic and collectivistic values. Thorstad and Wolff (2017) operationalised future-sightedness as the usage frequency of temporal expressions (such as *tomorrow, today, next year,* etc) on twitter and used it to predict inter-temporal choice and risk taking behaviour. However, caution must be taken regarding the assumptions behind selection criterion. For example, Greenfield (2013) reasonably assumes increasing usage frequency of *choose* suggests rising individualistic value because "freedom of choice is, by definition, a defining attribute of individualism". However, rising usage of *choose* after 1980s may be largely driven by its newly acquired meaning in technology contexts (signified by words such as *copy, command, mouse, click,* etc) rather than the rising individualistic values (figure 1.4). Since many polysemous words like *choose* are used across multiple different contexts, their relationship with the phenomenon of interest can be more obscure than we may realise. One should be aware that the assumptions based on which language is analysed may not always hold. When examining important assumptions is not possible, conclusions must be drawn with extra discretion (Pechenick, Danforth, & Dodds, 2015). For example, a good practise if to text whether results from analysing multiple corpora converge to the same conclusion.

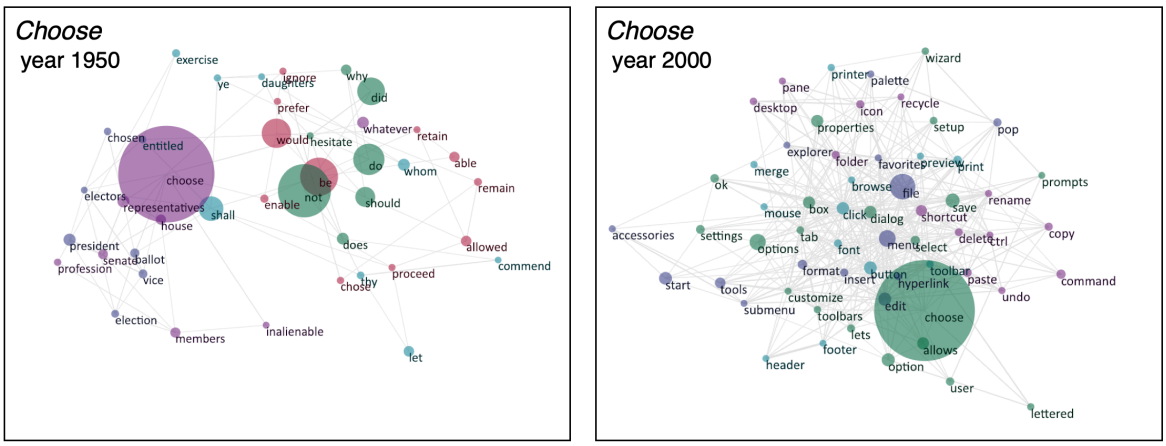


Figure 1.4. Both figures are produced from Macroscope (Li, Engelthaler, Siew & Hills, 2019). The left shows contextual network of choose in year 1950; the right shows contextual network of choose in year 2000. The nodes represent words that appear in the same context as choose. Edges are between words if their co-occurrence frequency exceeded a predetermined threshold. The size of nodes is proportional to their usage frequency. The colours represent the community structure of nodes in the network and each community is represented with a different colour.

Second, theory should be used to guide decisions on what aspects of corpus meaning should be analysed. Corpus meaning can be interpreted in light of its various linguistic properties. Some of these properties could be more suitable to answer a given question than others. For example, word frequency of particular words was found to be a good predictor of various psychological constructs such as depression (Eichstaed et al, 2015, Eichstaed et al, 2018), personality (Yarkoni, 2010; Schwartz et al, 2013), and sense of self (Tausczik & Pennebaker, 2010). Large-scale change in language in terms of concreteness and sentiment was found to correlate with social and demographic change (Hills & Adelman, 2015; Hills, Proto, & Srgoi, 2015). Semantic similarity between words for occupations – e.g., nurse, lawyer – and words representing women – e.g., *she, her* – quantifies gender bias and stereotypes of past 100 years (Garg, Schiebinger, Jurafsky & Zou, 2017). One can certainly analyse as many linguistic properties as one could to better understand the research questions and explore potential answers. However, selectively presenting analysis that converge to the desired results leads to data dredging (Coveney, Dougherty & Highfield, 2016). Therefore, all analyses should be evaluated in terms of their underlying assumptions, relevancy to the research questions and why they diverge/converge with each other.

I acknowledge that my approach to meaning of both word and corpus does not represent a complete picture. However, a holistic representation of word meaning can overwhelm investigators with too many details. As Herbert Simon once put it "A wealth of information creates a poverty of attention" (Simon, 1962), sometimes a subset of word meaning offers more insights when they are carefully chosen.

1.5 Research Questions

Words attain their meaning through experience. In turn, words we use in our daily life reflect who we are, what we feel and the social relationships we are in. This insight is neither new nor surprising. Language is the most common way for people to communicate their internal thoughts and emotions to others. Psychologists have been using it as an important medium to understand human minds. This attempt dates back to the earliest days of psychology: Freud (1901) wrote about parapraxes (slip of tongue) as discrepancy between what people actually said and what they intended to say. Therefore, such discrepancy indicates an

unconscious intention prevented from being expressed explicitly due to intrapsychic conflict. Rorschach (1921) developed projective tests to detect people's thoughts and motives from the way they described ambiguous inkblots. Similarly, Morgan and Murray (1935), using the thematic apperception test, found that the narratives people told in response to ambiguous pictures reveals important clues to their needs for affiliation. In all cases, trained raters read transcripts and encode the words and phrases according to dimensions the researchers were studying. In recent decades, research turns to use quantitative information of language such as like linguistic statistics instead of subjective interpretation. With the unprecedented size of many corpora (such as Google Ngram Corpus that consists of 6% of all published books over the past 4 centuries) that recently became available, investigation into historical sociocultural dynamics (e.g. Greenfield, 2013; Uz, 2014; Hills, Proto, & Srgoi, 2015) and language evolution (e.g. Hamilton, Leskovec, & Jurafsky, 2016; Xu & Kemp, 2015, Petersen, Tenenbaum, Havlin, Stanley, & Perc, 2012) has never been more approachable.

This thesis explores the potentials of language analysis in uncovering human minds. It starts with mapping words given by individuals to their affective states (chapter 2). We found that aggregated valence score across all words one had produced is a valid measure of general affect. It suggests individuals' psychological states can be reconstructed from the language they produced. In the next chapter, using 20 years of news articles published in the New York Times we studied the language around 56 immigrant groups in the U.S and explained why some groups are more favoured/feared than others (chapter 3). We quantified the immigrant-related language in relation to valence, concreteness (a proxy for social distance) and topics. We found language produced in public sphere can be used to infer the psychological representation of immigrant groups. It suggests that the notion of *words reflects minds* works at both individual and culture level. If such analysis is possible for every historical period, useful insights on cultural change may emerge. Given that Google Ngram Books Corpus contains word co-occurrence information over the past 4 centuries, we leveraged the abundance of this dataset and analysed semantic history of risk (chapter 4). In chapter 5, we present a linguistic tool we named the Macroscope that examines historical language structure. It makes most analysis we did for *risk* in chapter 4 and for immigrants in chapter 3 easily accessible for the top 50,000 most frequently-used English words. It takes one word as input, and generates a series of analysis as outputs including synonym analysis, contextual structure, semantic drift, historical trend of contextual sentiment, etc. Although each project is motivated by its own research questions, a common thread runs through these discussions: how language, when

analysed in different scales, informs individual minds, social attitudes, cultural change and language evolution.

Chapter 2 : Words to Individual Emotions

The Emotional Recall Task:
Juxtaposing Recall and Recognition-Based Affect Scales

Existing affect scales typically involve recognition of emotions from a predetermined emotion checklist. However, a recognition-based checklist may fail to capture sufficient breadth and specificity of individual's emotional experiences and may miss emotions that frequently come to mind. To address these issues, we present an affect scale based on recalled emotions. We asked participants to produce 10 words that best described their emotions over the past month and then to rate each emotion for how often it was experienced. We show that this task, the Emotional Recall Task (ERT), is strongly correlated with PANAS, Scales of Psychological Well-being, Satisfaction with Life Scale, the Office of National Statistics personal well-being measure, Depression Anxiety and Stress Scales, and the Beck Depression Inventory. We further show that the Emotional Recall Task captures the breadth and specificity of emotions that are not available in other scales but that are nonetheless commonly reported as experienced emotions. In addition, we show that the emotional fluency task is valid in a test-retest paradigm and that it can be reliably measured using paper and pencil. In sum, the emotional recall task supports recognition-based scales, but also offers a new direction for understanding differences in recalled and recognized emotions.

2.1 Introduction

> *"How people recall and estimate their moods is an important component of people's self-concepts and how they conceptualize their lives"*
>
> *(p. 292, Thomas & Diener, 1990).*

New affect scales often originate when limitations are identified in existing affect scales (Watson & Clark, 1998; Lucas, Diener, & Larsen, 2003; McDowell, & Praught, 1982; Thompson, 2007). Because all existing affect scales are recognition-based, previously identified limitations have often involved complaints that the list of terms on which participants base their emotional judgements "do not capture the range of people's experienced emotions" (Diener et al., 2009). In other words, the emotions that people experience are not those on the

recognition scale. Recognition scales require that people reinterpret their emotions in relation to emotions they may not have experienced or that may not readily come to mind in day-to-day experience. A scale based on recalled emotions might be a better indicator of people's affect across a broad range of emotions. Moreover, such a scale, by revealing where it is not predictive of people's recognized emotions, would offer insight into how emotions are accessed and the dimensionality of recalled versus recognized emotions. In this article, we introduce a recall-based affect scale, the Emotional Recall Task, and compare it with a number of currently popular recognition-based affect scales. Before introducing the Emotional Recall Task, we first briefly discuss the need for a recall-based emotional scale motivated by the history of research on emotional dimensionality. We then explain the potential differences in the memory literature between recall and recognition as they apply to emotions.

The specificity and breadth of emotional dimensions

A brief historical overview of the many approaches to dimensionalizing emotional experience shows two things: this is long-standing topic and there has been little historical consensus on exactly what and how many dimensions are important. The history of speculations about human emotions dates back to at least Aristotle's *Nicomachean Ethics* (Broadie & Rowe, 2002), which lists 11 different emotions, including 'pity' and 'emulation' (the act of copying another individual's behaviour). Darwin (1872), taking an evolutionary approach, attempted to classify emotions in relation to their adaptive value, and in addition to high and low valence emotions, included such dimensions as 'surprise', 'meditation' and 'shyness'. Looking across cultures, Ekman (1992) proposed a set of 'natural kinds' for emotions, similar to that of Darwin's, which included *anger, fear, disgust, sadness, happiness* and *surprise*. Wundt (1905) proposed that emotions largely fell along three dimensions: valence, arousal, and tension. Osgood, Suci, and Tannenbaum (1957) found further support for a similar three dimensions of meaning (evaluation-pleasantness, potency-control, and activity-arousal) using what is now called the semantic differential, which used a factor analyses of a large number of scales evaluating people's responses to various items. Russel (1980) proposed a circumflex model of emotion that suggests emotions are distributed in a two-dimensional space, with arousal and valence as independent dimensions. A more recent but similar approach based on principal components analysis found evidence for a fourth emotional dimension, unpredictability (Fontaine, Scherer, Roesch, & Ellsworth, 2007).

The discrepancies and agreements across this diversity of emotional primitives potentially stem from a number of sources. One key source is emotional granularity (Tugade,

Fredrickson, Barrett, 2004). Emotional granularity refers to an individual's ability to discriminate between different emotions. For example, a person with high (as opposed to low) emotional granularity would tend to express their emotions using more distinct words, like 'exuberant' (instead of 'happy'). A key individual difference identified in previous work is that people with less emotional granularity are more likely to focus on valence and may simply report degree of positivity or negativity, such as "very happy" (see Russell, 2003; Russell & Barrett, 1999). In other words, people differ in their emotional complexity (Lindquist & Barrett, 2008) and this may further indicate that they differ in their emotional dimensionality (Barrett, 2006). The apparent complexity of emotional dimensionality and our difficulties in establishing its universality may largely reflect individual differences in the way people experience affective states.

If people experience emotional dimensionality in different ways, this throws existing affective measurement scales into question. This is because the most popular approach to measuring emotions is to ask people about their ability to recognize how much they felt each of a set of emotions provided on a pre-determined checklist. Such recognition-based scales make two overarching assumptions. The first is that people will be able to identify their own emotions in relation to the words provided in the checklist. This we call the assumption of emotional *specificity*. The second is that the checklist will adequately cover a person's experience of emotions. This we call emotional *breadth*.

To put the ideas of emotional specificity and breadth in context, let us consider what is arguably the most widely used recognition-based checklist, the Positive and Negative Affect Schedule (PANAS) (for review see: Diener et al., 2010). The original article describing PANAS (Watson, Clark, & Tellegen, 1988) currently has more than 7500 citations as reported by Google Scholar. Because PANAS presents emotional stimuli, it necessarily frames respondents' emotional experiences in relation to emotions which may be more or less specific to the emotions respondents actually felt (e.g., Diener et al., 2009). PANAS focuses on a closed set of words, some of which are not generally considered as emotions (*strong*, *alert*, *inspired*, *determined*, and *active*), while common emotion words (*happy* and *sad*) are excluded. Four of the terms in PANAS focus on anxiety, and there are few low-arousal terms (Diener et al., 2009). Thus, PANAS's breadth is potentially narrower than the full emotional range that respondents experience.

Though PANAS is only one example, its potential problems of breadth and specificity are likely to be common to recognition-based scales more generally. Moreover, it may also suffer from order and priming effects (e.g., Hansen & Schantz, 1995; Wang, Busemeyer,

Atmanspacher, & Pothos, 2013). For example, being reminded of a forgotten emotion may make that emotion more salient than it otherwise would be in day-to-day experience.

One way to overcome these problems is to allow individuals to freely recall emotions they have recently experienced (e.g., in the last month). Because emotional terms are highly salient in free recall tasks (Altarriba & Bauer, 2004), the experience of an emotion may be easily recalled. Moreover, the recollection of emotional memories in a free recall task may be a better indicator of general emotional states and well-being than recognition-based scales because they reflect the emotional pathways laid down in the associative memory network (Bower, 1981), which plays a substantial role in the recollection of experience.

Recalled versus recognized emotions

Memory can be divided up into an effortful recollection-based process (recall) and less effortful familiarity-based process (recognition, see Raaijmakers & Shiffrin, 1981). Recall is the process of retrieving the details linked with a previous experience, while recognition is the process of identifying whether or not details presented to mind are present in memory. A principal difference between recall and recognition is therefore the retrieval stage of memory, which is not present in recognition-based scales (Anderson & Bower, 1972; Bahrick, 1970; Estes & DaPolito, 1967; Kintsch, 1970). In addition, several clinical studies have described cases where individuals have intact recognition memory but impaired recall memory, or vice versa, which suggest these processes may be controlled by different areas of the brain (Hanley, Davies, Downes & Mayes, 1994; Delbecq-Derouesne, Beauvois & Shallice, 1990).

The distinction between recognition and recall is therefore based on cognitive and neural differences and this may influence the kinds of emotions that come to people's minds in day-to-day experience and therefore their responses in different affect paradigms. For example, Tulving and Pearlstone (1966) observed that more memories may be available by recognition than by recall. At first glance, this appears to be a benefit to recognition-based scales. But this potentially comes with a cost of overlooking emotions that more frequently come to mind and of accurately assessing the frequency of recognized emotions. The availability heuristic refers to the well-documented observation that people often use the ease with which memories come to mind as indicators of their frequency and probability of occurrence (Tversky & Kahneman, 1973; Schwarz, Bless, & Bohner, 1991). As such, emotions that come to mind easily are likely to be those most frequently experienced. In addition, previous studies have found that the effort involved in recall may be a better cue to the accuracy

of a memory. For example, Robinson & Johnson (1996) found that a recall-based measure of eyewitness memory led to a better confidence-accuracy correlation, indicating that recall provided additional information that was lost in assessments based only on recognition (see also Koriat & Goldsmith, 1996). This is potentially a problem for recognition-based scales.

The challenge we set forth here is to create a recall-based affect scale and compare it with existing recognition-based scales. Van Rensbergen, Kuppens, Storms and De Deyne (2015) demonstrated the ability to use a recall-based scale in assessing the Big Five personality traits. Their experiment asked participants to describe their personality using ten adjectives. Participants' personality scores were then obtained from the average correspondence between these adjectives and the Big Five personality factors. In Study 1, we follow Van Rensbergen et al.'s (2015) lead by investigating an emotional recall task that asks participants to recall and rate recent emotions. We then compare this against existing recognition-based metrics. In Study 2, we present a comparison of test-retest reliability, showing that a recall-based measure, the ERT, is on par with existing recognition-based scales. Finally, in Study 3, we present a paper and pencil version of the ERT, demonstrating that a recall-based measure—which is content neutral given that participants can produce whatever emotions they choose—can be easily administered, making it ideal for assessments across languages, ages, and cultures.

2.2 Study 1: Comparison and validation of recall and recognition-based scales

Study 1 compares the ERT with several standard recognition-based scales. The central goal in this study is to evaluate the external validity of a recall-based measure, the ERT. The ERT encourages people to actively search their memory for emotions they have experienced. Participants are required to first produce 10 words to best describe their feelings over a recent period of time. Next, they rate each of these words on a 100-point scale to indicate how frequently they have experienced these feelings.

Methods

Participants. 130 participants were recruited from Amazon Mechanical Turk. They are based in the United States and reported as native English speakers. We excluded 4 participants from the analysis because they failed to follow instructions. This left us with 126 participants (male = 57, female = 69).

Procedure. The questionnaire was administered on Qualtrics. Following the consent form, participants were taken to a webpage and provided with the following instruction: "Please list 10 words that best describe your feelings in the past month". After entering 10

words, a second page appeared representing the 10 words the participant produced in a randomized order with an instruction to "indicate how frequently you have experienced each of these emotions on the slider below". The slider ranged from 0 (not often at all) to 100 (very often). All participants filled out the ERT first to avoid being primed with words from other scales. Following this, they were randomly presented with each of the following scales:

The Positive and Negative Affect Schedule (PANAS, Watson et al., 1988) consists of two 10-item mood scales. It was developed to provide a brief measure of positive and negative affect. The 20 PANAS items were derived from a principal component analysis of Zevon and Tellegen's (1982) 60-item mood checklist. Respondents are asked to rate the extent they experienced each emotion within a specific time frame, with reference to a 5-point scale that ranges from 'very slightly or not at all' to 'very much'. Different time frames (e.g., "right now", "today", "during the past few days", "during the past week", "during the past few weeks", "during the past year", "in general") have been used with the PANAS. In the present study we set time frame to "during the past month".

The Ryff Scales of Psychological Well-Being (SPWB, Ryff & Keyes, 1995) is a theoretically grounded instrument that specifically focuses on measuring multiple facets of psychological well-being. These facets include the following: autonomy, environmental mastery, personal growth, positive relations with others, purpose in life, and self-acceptance. Individuals respond to various statements and indicate on a 6-point Likert scale on how true each statement is of them. Higher scores on each scale indicate greater well-being on that dimension. We used the 18-item version in the current study.

The Diener Satisfaction with Life Scale (SWLS, Diener, Emmons, Larsen, & Griffin, 1985) is a short 5-item instrument designed to measure global cognitive judgments of satisfaction with one's life as a whole. The scale does not assess satisfaction with life domains such as health or finances but allows subjects to integrate and weight these domains in whatever way they choose.

The ONS-4 was developed by the Office for National Statistics of UK to assess personal well-being using 4 measures that capture 3 types of well-being: evaluative, eudemonic and experience (*Tabor & Stockley, 2018*). These measures ask people to evaluate the overall life satisfaction, worthiness of things they do, happiness, and anxiety. It was first added to the Annual Population Survey (APS) in April 2011 and has been used in many surveys across the UK .

The Depression Anxiety and Stress Scales (DASS-21, Lovibond & Lovibond, 1995) consists of three 7-item self-report scales that measure depression, anxiety, and stress correspondingly. Each items was rated on a 4-point scale.

The Beck Depression Inventory (BDI, Beck, Steer, & Brown, 1996) measures severity of depression in normal and psychiatric populations. The questionnaire was developed from clinical observations of attitudes and symptoms occurring frequently among depressed psychiatric patients and infrequently in non-depressed psychiatric patients. The questionnaire contains 21 questions, each ranging on a scale from 0 to 3.

Construction of ERT scale. In Study 1 we use the valence norms of Warriner, Kuperman and Brysbaert (2013) to compute the valence for each word. The Warriner et al. norms are an extended version of Bradley and Lang's (1999) Affective Norm for English Words (ANEW), providing ratings of valence for almost 14,000 English words. Each word was rated by around 20 participants on a scale from 1 (unpleasant) to 9 (pleasant). This database allows us to transform a list of emotions collected from each participant into a vector of valence. The overall affective state of each participant can then be calculated using the formula below:

$$\mathrm{V} = \frac{1}{10}\sum_{i=1}^{10}(V_i - 5) \times R_i$$

where V denotes overall affective state of an individual in terms of valence. This is the ERT score. R represents the self-reported frequency of the i^{th} feeling. V_i denotes the respective valence rating of the i^{th} feeling from the Warriner et al. valence norms.

Over all participants, there were 139 words that cannot be transformed into valence ratings because those words are not included in the norm database (Warriner et al., 2013). To tackle this issue, we used Word2Vec (Mikolov, Chen, Corrado, Dean, 2013), a language model that provides semantic similarity between two words, to replace non-existent words with the most proximal words that exist in the extended version of ANEW. This allowed us to use all of the words each participant produces. As we show in the final section this step can be eliminated by having participants rate their own words.

Results

Participants produced 466 unique words and 64% were mentioned only once. Our analysis shows participants tended to first recall emotions they experienced more frequently, with less frequent words produced later in the sequence (Fig. 2.1a). Emotions produced earlier in the recall sequence were also produced faster than later words (Fig.2.1b). Figure 2.1c shows that the valence of emotion words are bimodally distributed, suggesting that people experience more non-neutral emotions than neutral emotions (Fig. 2.1c).

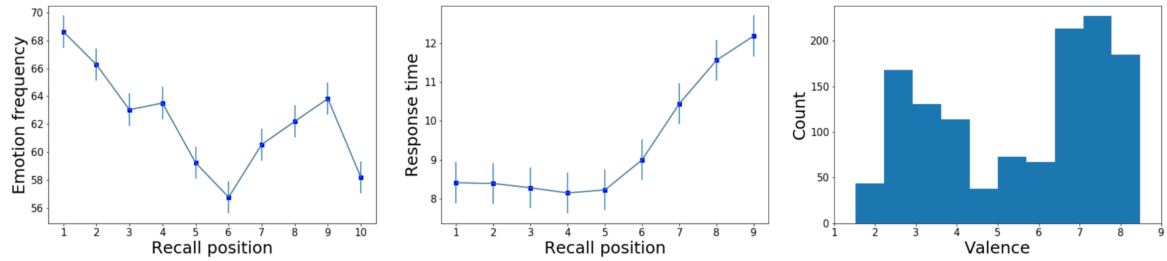


Figure 2.1. Statistics on words produced in the ERT. The error bars represent standard errors. a) the average frequency of experiencing reported ERT emotions in each recall position. b) the average time (in seconds) spent on generating ERT emotion words in each recall position. c) distribution of valence values for all terms produced in the ERT.

How is the ERT different from the PANAS in terms of emotional breadth and emotional specificity? Figure 2.2a shows that few PANAS terms are among the most popular words that people used when describing their past feelings. Only 1 PANAS term (*excited*) appeared among the top 10 most frequently recalled emotions. This raises concerns that participants may not have identified their past feelings in relation to many of the terms in the PANAS.

Our results also present quantitative evidence that PANAS suffers from issues of emotion breadth. Figure 2.2b compares distribution of PANAS terms and the ERT terms on the affect space of valence and arousal. It shows that ERT terms distribute across the entire arousal space while the PANAS contains no low arousal emotion terms. Moreover, although both scales cover the extreme ends of valence space, PANAS has few neutral terms.

Convergent validity. A good emotion scale should be able to predict related constructs. We first analysed the relation between the ERT and PANAS. The pairwise correlation coefficient of PA, NA and the ERT can be found in Table 2.1 alongside other scales (which we discuss in the subsequent section). Consistent with previous studies of PANAS, we found the PA and NA component are independent of each other (r = -.14, p = -0.12). The ERT correlate with both PA and NA (r = 0.56, $p < 0.001$ for PA and r = -0.59, $p < 0.001$ for NA).

Table 2.1 Correlation table between all measures

		1	2	3	4	5	6	7	8	9	10	11	12	13
ERT	1. ERT													
PNAS	2. PA	0.57***												
	3. NA	-0.58***	-0.14											
SWLS	4. SWLS	0.65***	0.57***	-0.26**										
SPWB	5. SPWB	0.58***	0.43***	-0.57***	0.54***									
ONS	6. Life satisfaction	0.75***	0.64***	-0.43***	0.87***	0.61***								
	7. Life worthiness	0.69***	0.61***	-0.39***	0.72***	0.68***	0.83***							
	8. Happiness	0.7***	0.53***	-0.47***	0.63***	0.56***	0.78***	0.71***						
	9. Anxiety	-0.52***	-0.17	0.73***	-0.25**	-0.53***	-0.39***	-0.36***	-0.52***					
BDI	10. Depression	-0.69***	-0.36***	0.75***	-0.44***	-0.64***	-0.58***	-0.55***	-0.58***	0.67***				
DASS	11. Depression	-0.69***	-0.38***	0.68***	-0.45***	-0.69***	-0.62***	-0.66***	-0.65***	0.63***	0.87***			
	12. Anxiety	-0.39***	-0.03	0.65***	-0.05	-0.47***	-0.2*	-0.24**	-0.27**	0.58***	0.64***	0.72***		
	13. Stress	-0.53***	-0.21*	0.71***	-0.29***	-0.54***	-0.47***	-0.48***	-0.53***	0.63***	0.75***	0.83***	0.81***	

Note: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

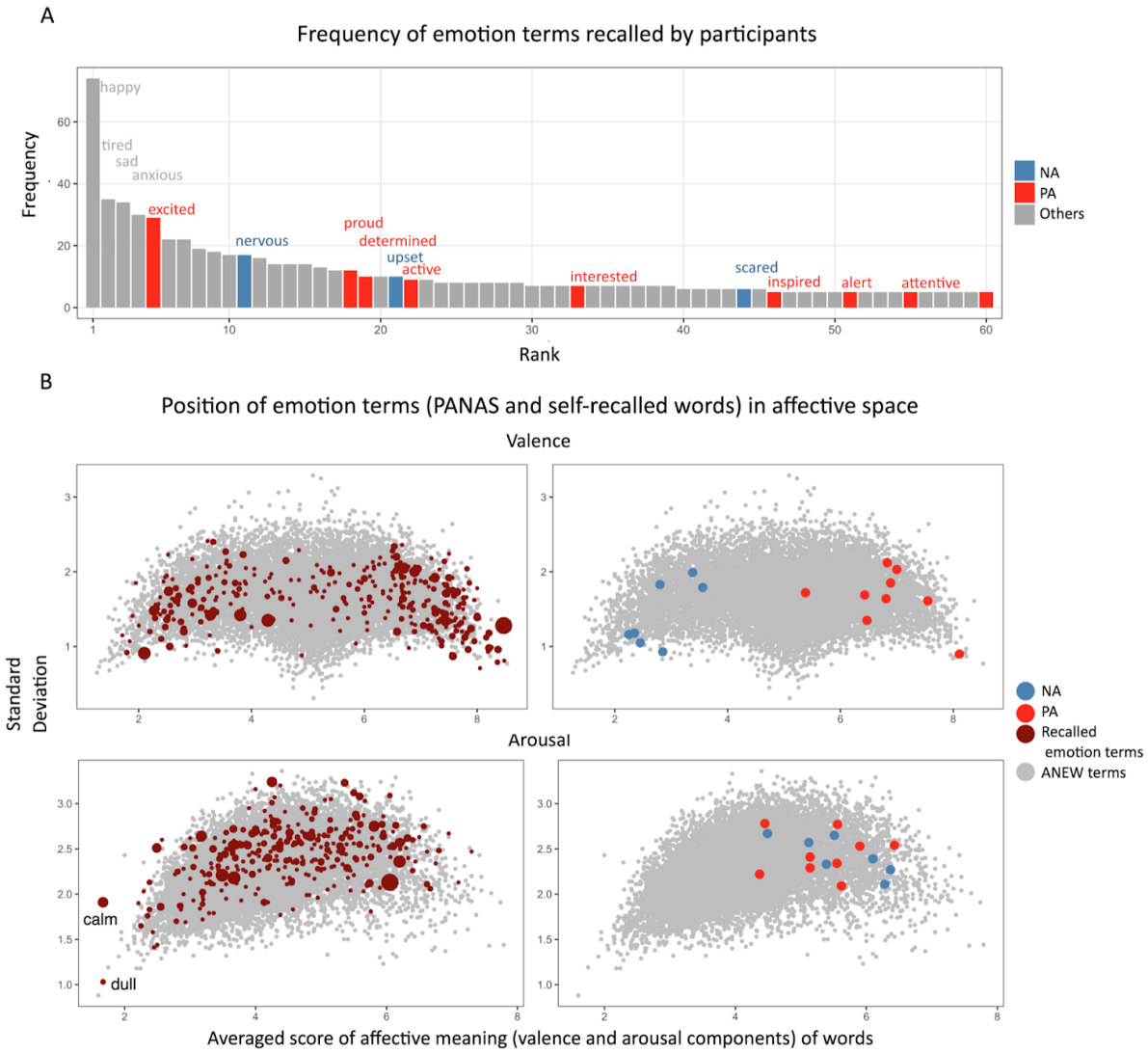


Figure 2.2. Emotional breath and specificity of the ERT and the PANAS. A) shows the frequency of words recalled in the ERT and where the PANAS words are located in the ERT frequency ranking (highlighted in red and blue respectively for positive affect and negative affect). B) shows where the PANAS terms and the ERT terms are located along the dimensions of valence and arousal. The x-axis is the mean valence or arousal rating and the y-axis is the standard deviation of these ratings. Higher standard deviation indicates larger degree of disagreement amongst those rating the words in the norms. Each grey dot represents one word from the existing affective norm database (Warriner et al., 2013).

We further explored the discrepancy between the ERT and the PANAS by examining participants whose emotional states were inconsistent between the two measures. Figure 2.3A shows how participants' ERT scores are related to the PANAS. Several individuals are particularly noteworthy. In the ERT, participant 15 (ID number = 15) generated a number of negative emotion terms, and no positive terms, and reported experiencing each of the negative terms with high frequency (Fig.2.3 B2). Yet this participant reported extremely low negative affect in the PANAS scale (Fig.2.3A left). Similarly, participant 72 recalled 8 positive emotions,

1 neutral emotion and 1 negative emotion (Fig.2.3 B4). But the same participant's PANAS score suggests the participant experienced little positive affect. Participants 75 and 66 (Fig.2.3 B1 and B3) show similar discrepancies between recalled and recognized emotions.

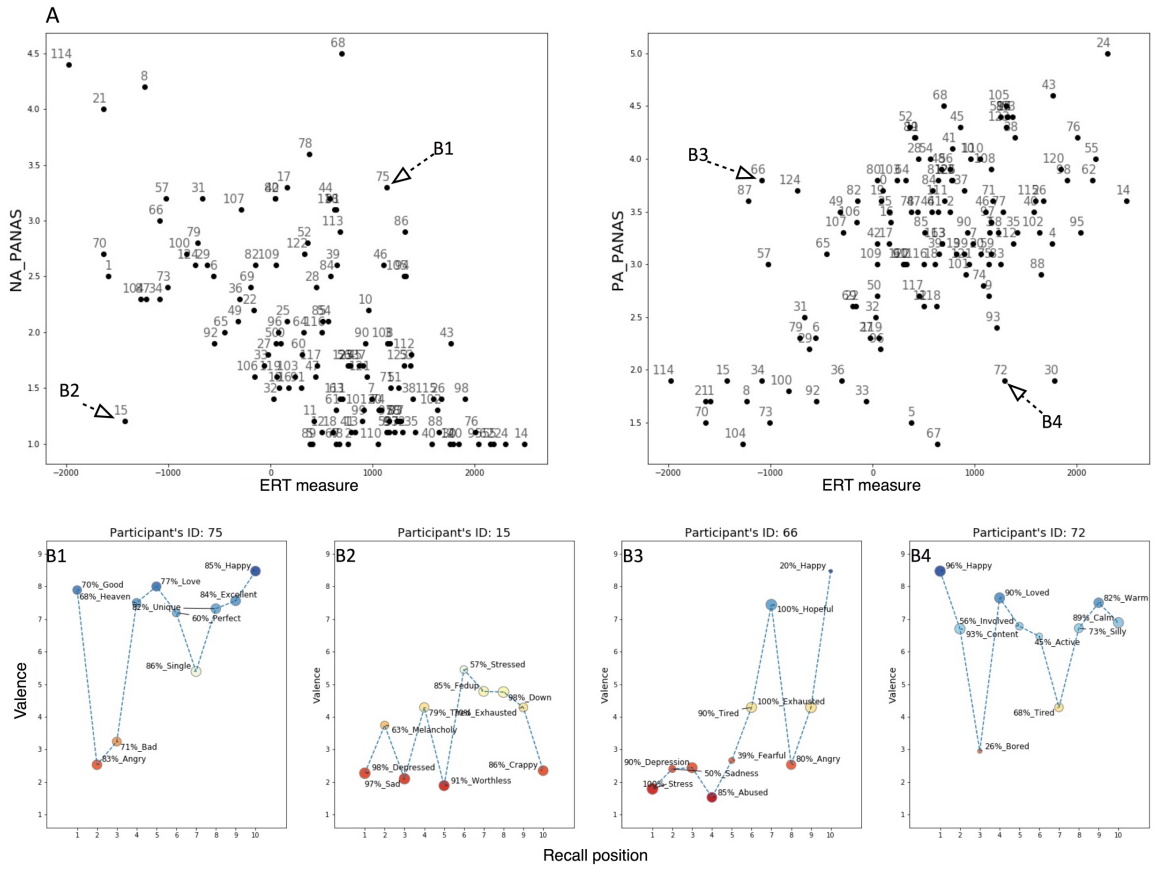


Figure 2.3. Discrepancy between the ERT measure of emotion and the PANAS. Figure 2.3A shows correlation between ERT measures and NA and PA of the PANAS. Figure 2.3B1—2.3B4 shows the sequence of 10 words produced by the 4 participants identified in A and also provides their frequency (in %) next to each entry. Color shows word valence (blue = positive, red = negative) and dot size corresponds to frequency.

Correlation with other related constructs. To further test the validity of the ERT, we compared it with the PANAS on how well it predicted related constructs. Table 2.1 shows that the ERT performs at least as well as the PANAS in predicting the 3 wellbeing-related constructs (Diener, Ryff and ONS4), and 2 depression measures (BDI and DASS). In particular, the ERT has higher correlations for all additional scales than does PA for the PANAS scale. On the other hand, the NA of PANAS performs better in predicting ONS anxiety, BDI Depression, and DASS Depression, Anxiety, and Stress. This may not be surprising since, as noted above, 4 out of 10 terms in the NA portion of the PANAS scale are anxiety related (Diener et al., 2009). Nonetheless, though the correlations are marginally better or worse in many cases, the correlations are generally high across all scales, indicating that the ERT is well-positioned with respect to existing scales.

Are 10 emotion terms sufficient? To test whether 10 words is sufficient to capture emotion experience in the ERT, we performed a sensitivity analysis to show how correlational strength between the ERT and other constructs change in relation to the number of emotion terms included. Figure 2.4 shows that the correlation generally improves across the 10 words. This improvement has a diminishing marginal return: the improvement plateaus at between roughly 4 and 10 words depending on the specific scale one uses for comparison.

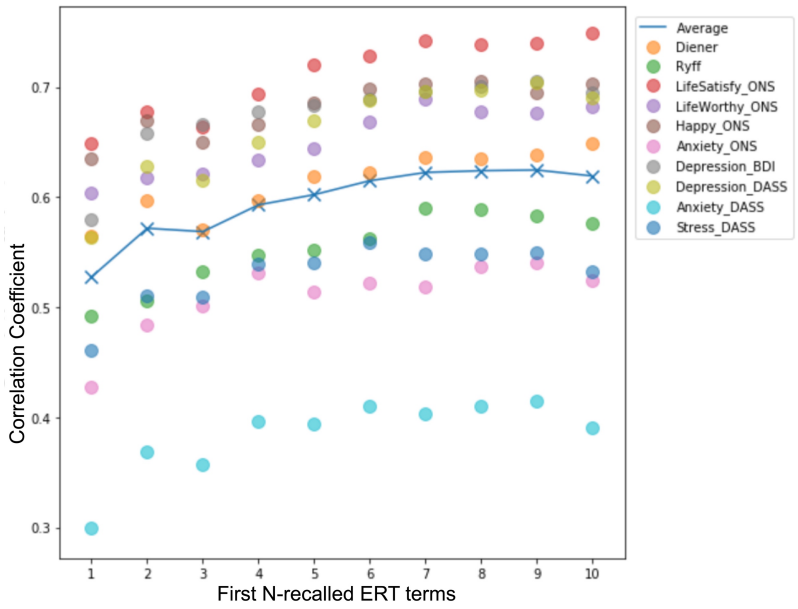


Figure 2.4. Sensitivity analysis between the ERT measure and other constructs in relation to increasing number of the ERT words included (in recall order).

2.3 Study 2: Test-retest reliability

In this study, we examine and compare the test-retest reliability of the ERT against other scales.

Methods

Participants. The ERT scale was given twice to a group of 119 undergraduate students from the University of Warwick. Seven failed to complete the first or second test. The remaining 112 students completed both test and retest and are included in the analysis below. Students were compensated with course credit. The test-retest scale was approved by the University of Warwick's ethics approval board. Participants reported as female in 90 cases (80.35%) and as male in 22 cases (19.64%). The mean age of participants was 19 years (M=19.08; SD=1.08) ranging from 18 to 26 years.

Measures. Participants in the test-retest study filled out the ERT, the PANAS, Diener's SWLS, the ONS, and BDI-21. These are as described in the previous section.

Procedures. Participants were invited to participate in an online study where they would be provided with a set of survey questions twice, with the two occasions separated by at least two weeks. They were asked to provide a matching identifier in both tests that could be used to match responses for each individual between the two separate occasions. Participants received the link for the second survey 14-days after completing the first. All other details are as in Study 1 above.

Results

Table 2.2 presents the reliability results for each of the tests. Because only the ERT fulfilled the requirements of the Shapiro-Wilk test for normality, we use the non-parametric Spearman rank correlation for all tests. Table 2.2 shows that the various scales all have comparable reliability. Moreover, the scores from scales in prior work are aligned with those found here (e.g., PA, .58, $p < .05$ and NA, .48, $p < .05$, from Watson et al., 1988; SWLS, .82 from Diener, et. al., 1985; BDI, .67 from Fydrich, Dowdall, & Chambless, 1992). Note that the 95% confidence interval for each of the tests overlaps with the Spearman rank correlation for the ERT. The ERT yielded correlations on par with existing affect scales. We found that ONS happy and ONS anxiety has the smallest test-retest correlations because the questions were framed to ask one's emotion states "today" and therefore more sensitive to daily events. The ERT has larger test-retest correlation than the ONS happy/anxiety but lower than the PANAS. It may suggest that the recalled emotion is more sensitive to everyday events.

Table 2.2 Test-retest reliability correlations between affect scales

Affect Scales	**Correlation Index (Spearman Rank, *n* = 108)**	**95% CI**
ERT	.42***	.25, .57
PANAS Positive Affect	.57***	.43, .69
PANAS Negative Affect	.53***	.38, .65
Satisfaction With Life Scale	.69***	.52, .74
ONS Life Satisfaction	.55***	.40, .67
ONS Life Worthiness	.40***	.23, .55
ONS Happy	.26***	.07, .43
ONS Anxiety	.38***	.20, .53
Beck Depression Inventory	.56***	.42, .68

Note: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

2.4 Study 3: Paper and pencil version of the ERT

One issue with the ERT is that it uses valence norms to compute the valence for each recalled emotion. This has two problems. First, this assumes that different people share the same valence ratings for words. As shown in figure 2.2b, this assumption is false, since words have non-zero standard deviations in their valence and arousal ratings. *Workaholic* may be a pleasant term to someone who enjoys his work, but an unpleasant term to others. In addition, referring to an English database would limit the ERT's generalizability to other cultural and social groups where affective norms are not available. Second, participants may produce words that are not in the affective norms. Although we use machine learning above to replace those words with semantically similar words, this introduces additional computation and possibly error.

To solve these problems, Study 3 examines a method to retrieve valence by asking participants to rate the valence of the words they produce. This allows us to compare performance between the norms-based Emotional Recall Task (henceforth ERT 1.0) and the 'paper-and-pencil' version of the Emotional Recall Task (henceforth ERT 2.0). We use the term paper-and-pencil to indicate the ease with which the test is given and scored, but it nonetheless can (and did in the present study) take place on a computer. In this modified version of the ERT, participants produce ten emotions they have felt in the past month, and then they rate these emotions for how often they have felt each of them. Finally, they rate each emotion for its valence on a scale from 1 to 10.

Participants. The tasks were presented to 200 native English speakers recruited from Amazon Mechanical Turk (MTurk). Four participants were removed because they failed to follow instructions.

Procedure. The procedure is exactly the same as study 1 except after reporting 10 emotion words, participants were additionally required to evaluate the valence of each word on a scale from 1 to 9. In a departure from Study 1 and 2, we instructed participants to use only one word to describe their feelings since Warriner's affective norms do not include any 2-grams. However, participants may often find it easier to use two or more words to describe their feelings. For example, we have observed responses such as *out of control* and *in love*, among others. These emotion terms could cause problem for the ERT 1.0 but not for the ERT 2.0.

Results

Table 2.3 shows that results from the ERT 1.0 and the ERT 2.0 are highly correlated ($r = 0.97$, $p<0.001$). As in Study 1, both versions of the ERT performs at least equally well as PANAS in predicting measures of well-being and depression, while the negative component of PANAS better predicts anxiety and stress. This strongly suggests that the Warriner et al. norms can be used accurately to capture emotional affect in the ERT, but also indicates that the ERT can be used reliably without the need for using valence norms or machine-learning. Indeed, the ERT 2.0 can be easily computed on the back of an envelope.

Table 2.3 Correlations between ERT 1.0, ERT 2.0, and all related constructs

		1	2	3	4	5	6	7	8	9	10	11	12	13
ERT	1. ERT v1													
	2. ERT v2	0.97***												
PANAS	3. PA	0.75***	0.74***											
	4. NA	-0.69***	-0.7***	-0.51***										
SWLS	5. SWLS	0.73***	0.72***	0.71***	-0.54***									
SPWB	6. SPWB	0.72***	0.69***	0.71***	-0.62***	0.78***								
ONS	7. Life satisfaction	0.77***	0.75***	0.73***	-0.56***	0.92***	0.82***							
	8. Life worthiness	0.73***	0.72***	0.73***	-0.53***	0.83***	0.84***	0.9***						
	9. Happiness	0.82***	0.8***	0.73***	-0.6***	0.79***	0.8***	0.86***	0.84***					
	10. Anxiety	-0.64***	-0.64***	-0.51***	0.73***	-0.49***	-0.6***	-0.53***	-0.52***	-0.6***				
BDI	11. Depression	-0.69***	-0.68***	-0.59***	0.66***	-0.69***	-0.75***	-0.72***	-0.73***	-0.75***	0.57***			
DASS	12. Depression	-0.72***	-0.7***	-0.61***	0.64***	-0.72***	-0.78***	-0.78***	-0.78***	-0.78***	0.58***	0.87***		
	13. Anxiety	-0.44***	-0.46***	-0.3***	0.5***	-0.41***	-0.46***	-0.43***	-0.45***	-0.45***	0.57***	0.64***	0.63***	
	14. Stress	-0.64***	-0.64***	-0.45***	0.72***	-0.52***	-0.58***	-0.53***	-0.55***	-0.6***	0.68***	0.71***	0.75***	0.72***

Notes: *$p<.05$, **$p<.01$, ***$p<.001$

2.5 Discussion.

We found that the Emotional Recall Task, by relying on recalled memory of emotional experience, captures affective states and correlates highly with other commonly used measures of well-being. In addition, both correlations (between PANAS and ERT) and individual case studies (Figure 2.3) suggest that the ERT captures different aspects of emotional experience from the PANAS, suggesting interesting areas for future research.

One important limitation to PANAS and other existing recognition-based affect scales is their limited generalizability. People across various cultural and social groups often have systematic differences in their experienced emotions. For example, *schadenfreude* is a German word (meaning deriving happiness from another's pain) for which there is no such word in English. Scollon, Diener, Oishi, and Biswas-Diener (2004) have identified numerous emotion terms important that are often not shared across cultures. Even within the same culture, the PANAS can be problematic when comparing scores across different social or age groups. Because PANAS lacks low-arousal terms, young people may score higher on positive affect than the elderly just because they are generally more sensation-seeking (Oishi, Schimmack, &

Colcombe, 2003). The elderly may score higher on pleasant terms such as 'contented' and 'peaceful', but these low-arousal feelings are not assessed by the PANAS. Therefore, a recognition-based approach to emotion may carry numerous assumptions that do not generalize across cultures or within a same culture because people vary from one to another in terms of what emotions were experienced and valued.

The ERT overcomes this problem by allowing all individuals to freely choose emotion terms that readily come to mind. This has additional advantages as well. At a macro level, the ERT scores can be used to make comparisons across individuals just like other recognition-based scales, while at a micro level, the ERT offers greater details on the breadth and specificity of an individual's emotions, how these emotions are searched, and why some individuals may perseverate on negative emotions while others do not. Most importantly, the comparison provided here offers a new approach to investigate the differences between recalled and recognized emotions.

In closing, our analyses suggest that the ERT is a reliable, valid, and content neutral means for measuring emotional recall. Relying on recall process, it caters to individuals' specific emotional experiences and avoids several disadvantages shared by all recognition-based scales, namely, emotion specificity and emotion breadth. As noted in the introduction, specificity and breadth have seen a long history of discussion that is unlikely to be resolved given people's individual differences. The ERT task offers a new approach to studying these differences.

Chapter 3 : Words to Social Attitudes

Quantifying Historical Change in Patterns of Immigrant Sentiment

Public perception towards immigrants is complex and multi-faceted. Some people celebrate economic development, innovation, and the cultural diversity immigrants bring to a country while others blame immigrants for erosions of national competitiveness, organized crime, and illegal immigration. Although psychological factors influencing these attitudes have been identified in the past, it remains unclear how attitudes of immigrant groups change over time and why hostility towards immigrants are directed towards some immigrant groups but not others. In the present study, we quantify historical change in language around 56 immigrant groups in the U.S from an immigrant corpus derived from newspaper articles over a 20-year period. This is quantified in relation to sentiment, concreteness (a proxy for social distance) and 15 topics derived from Latent Dirichlet Allocation identifying issues such as crime, terrorism, illegal status, books, religion, cuisine, and art. In support of prominent theories of outgroup prejudice and intergroup contact, positive sentiment is strongly correlated with concrete descriptions, with concrete language predicting future positivity but not vice versa. Positively viewed immigrants are also best associated with topics of positive sentiment, such as cuisine, movies, and art. Together, these suggest implications for policy aimed at reducting intergroup conflict and future research.

3. 1 Introduction

According to World Bank World Development Indicators (2017) there were approximately 250 million international immigrants worldwide in 2015. These migrants have consequences for economics, health, international conflict, and the political futures of nations most recently exemplified by Brexit and numerous world elections. What these migrant numbers fail to reflect is that most of the world's ethnic groups were at one time immigrants who have themselves suffered greater or lesser degrees of discrimination, assimilation, and outgroup bias. Immigrants are commonly perceived as untrustworthy outsiders (Cuddy, Fiske, Demoulin, & Leyens, 2000; Eckes, 2002; Cuddy et al., 2007; Peabody, 1985; Poppe, 2001; Alexander, Brewer and Herrmann, 1999), even though they bring innovation, skilled labor, investment, and rich cultural diversity (Borjas, 1990; Carens, 2013; Skeldon, 2014). Understanding the mechanisms that form and revise public perception of outgroups is central

to providing solutions that enhance outcomes for immigrants as well the populations into which they migrate.

Central to immigrant outcomes is the role of outgroup negativity, which is deeply rooted in basic human propensities for social categorical thinking (Allport, 1954, Brewer, 1979, Tajfel, 1982). *Ultimate attribution error,* for example, is the propensity for people to explain negative behaviour of others as dispositional properties of a categorically defined outgroup, but explaining other's positive behaviour as a result of idiosyncratic situational factors (Pettigrew, 1979). What is remarkable though is that outgroup status and sentiment is decidedly flexible. Laboratory analogues of group formation, often called *minimal group paradigms*, demonstrate that the minimum condition for intergroup bias is categorization into a group, even with arbitrary criteria for categorization, such as the preference for Kandinsky over Klee (Tajfel et al, 1971). Sherif's (1961) Robbers' Cave experiment demonstrated rising prejudice and hostility over a period of weeks towards outgroup members merely by assigning boys to arbitrary groups at a boys' camp. Sherif (1961) further showed that this outgroup status could be rapidly ameliorated through cooperative action towards a common goal.

What factors influence the reconceptualization of group boundaries in relation to immigrants? One of the most prominent and well-supported theories is *intergroup contact theory* (Allport, 1954): an effective way to resolve outgroup prejudice is through reducing social distance, via direct interaction. A meta-analysis of more than 500 studies of intergroup contact theory found that 94% of independent samples showed increased contact reduced prejudice (Pettigrew & Tropp, 2006). Generally speaking, social distance promotes dispositional inference and prejudice (Jones & Nisbett, 1972; for a review, see Gilbert, 1998). For example, intergroup contact plays a substantial role in explaining the rural-urban divide in immigrant perception, whereby rural populations with the least contact with immigrants tend to have higher outgroup negativity than urban populations (Fenelly & Federico, 2008). This represents a distinct description-experience gap (Hertwig & Erev, 2009), whereby learning from direct intergroup contact benefits outgroups but learning filtered through media descriptions perpetuates racial and gender biases, which are known, for example, to negatively influence machine learning algorithms trained on the same material (Bolukbasi, Chang, Zou, Saligrama, & Kalai, 2016; Caliskan, Bryson, & Narayanan, 2017).

Intergroup contact theory, however, is more nuanced than simple endorsement of contact. Contact must have pro-social qualities, such as equal status, cooperation, and social approval, all of which reduce perceived social distance (Allport, 1954). Despite threats of a clash of civilizations (Huntington, 1993), the growing pains of intergroup assimilation may

benefit from sufficient proximity to experience many of these positive factors. This may already be happening. A series of replications of Bogradus (1927) study of contact among divers social groups in the United States have found that among 30 ethnic groups all were perceived as less socially distant now than in the past (Parrillo & Donaghue, 2005).

What these previous studies do not provide is a comparative understanding of the social contexts that predict change in sentiment towards immigrants over cultural time. While excellent case studies from sociologists and psychologists have investigated social context and offered detailed analysis on how immigrant attitudes were shaped under specific social, economic, and political environments (Allport, 1954; Portes & Zhou, 1993; Portes & Sensenbrenner, 1993), they have predominantly focused on a few target groups and a few issues per study and only over relatively short intervals of time. Ideally what we would like to know is how perceptions towards a wide variety of immigrant groups have changed over cultural time as mediated by perceived contact in relation to a range of social contexts.

To examine these questions, we first analysed an "immigrant corpus" first at a macro, quantitative level by investigating two distinct aspects of attitudes towards immigrants—*sentiment* (or valence, taken to indicate positive or negative perceptions) and language *concreteness* (indicating direct experience, used to operationalize social distance). We hypothesize that perceived social distance towards an immigrant group predicts its future sentiment. This was done by analysing language around 56 immigrant groups from a corpus containing 20 years of news articles published in the *New York Times* (Sandhaus & Evan, 2008). We constructed an immigrant news corpus by selecting all articles that contain at least one appearance of *immigrant* or its variations. Articles mentioning immigrants of the same ethnicity were grouped together (referred to as *ethnic corpora* in the following text). Sentiment of an ethnic group was calculated by taking all the words from the corresponding ethnic corpora and computing their mean valence. We used the valence norms of Warriner, Kuperman and Brysbaert (2013). This is an extended version of Bradley and Lang's (1999) Affective Norm for English Words (ANEW). The use of concreteness to measure social distance is supported by construal level theory, which has shown that representing a person abstractly reflects perceived social distance (Trope & Libermann, 2010). For example, concrete language has been shown to fall off as one moves from describing family to friend to coworkers to foreigners (Snefjella, Bryor, & Kuperman, 2015). Concreteness and abstractness can be measured in language in a similar way to how sentiment is measured, by evaluating the perceived concreteness of individual words that make up the text (Hills & Adelman, 2015). We computed

language concreteness using the 40,000 word concreteness norms provided by Brybaert et al (2014).

To evaluate the context underlying and driving sentiment, we inferred immigrant-related topics by applying Latent Dirichlet Allocation (LDA; Blei, Ng, & Jordan, 2003), a topic modelling algorithm that identifies underlying patterns (or topics) that best explain corpus structure using Bayesian inference. This allows us to tease apart the underlying social contexts that may drive positive or negative sentiment and evaluate with respect to changes in immigrant sentiment over time.

3.2 Materials & Methods

Immigrant corpus from the New York Times Annotated Corpus. The language corpus we used is the New York Times Annotated Corpus (Evan, 2008). It contains nearly all articles (over 1.8 million) written and published by the *New York Times* between January 1981 to June 2007. To retrieve immigration-related news articles, we include all articles containing at least one appearance of the word 'immigrant' or its variations ('immigrants', 'immigration', 'immigrate'). This procedure rendered an 'immigrant corpus' containing 43,350 articles. Next, in order to examine language used on immigrant groups from different ethnic groups, we constructed ethnic corpora by selecting articles mentioning each ethnic group from the 'immigrant corpus' so that, for example, the Mexican corpus contains all articles mentioning 'Mexicans' as immigrants. In our study, we investigated 48 immigrant groups by their country of origin and 8 ethnic categories considered important components of the U.S society (such as African American and Latino) by Bogradus (1927) and Parrillo and Donoghue (2005). Immigrant groups were selected based on their population size reported in the American Community Survey (U.S. Department of Homeland Security, 2010): only those groups consisted more than 0.8% of the total population were included.

Corpus concreteness and valence. We computed concreteness of news articles using a recent data set of concreteness ratings for 40,000 English words (Brybaert et al., 2014). It was developed by taking the average ratings of words on a scale from 1 (abstract) to 5 (concrete) as taken from 30 participants, resulting in concreteness norms ranged from 1.04 ('essentialness') to 5 ('pitbull'). Concreteness of an article was computed by taking the mean concreteness rating of all words in that article. Similarly, we inferred word valence using affective norms of English words collected by Warriner et al. (2013). It is a database of nearly 14 thousand English words, all rated on a scale from 1 to 9. Each word was rated by 20 participants and the mean valence

rating of each word was used for this study. Concreteness and valence of a corpus was computed by averaging across ratings of articles contained in that corpus.

Topic model. LDA assumes there is a set of latent patterns (or topics) that explain and generate the structure of textual documents. It computes documents as a distribution of topics over document with topics themselves represented as distributions of words. LDA was trained on the immigrant corpus such that each documents was assigned a distribution of topics and each topic was made up a distribution of words.[1]

For instance, a word vector of the form "dangerous illegal workers" in one document may be translated to "10 2 2", in which the last two words in that document were generated by topic 2, and the first word by topic 10. The same word can be assigned to different topics, allowing generic words to appear in multiple topics.

To make sense of topics, we examined the 10 most relevant words for each topic. We defined the relevance of term *w* to topic *k* (Sievert & Shirley, 2014) as:

$$\gamma(w, k|\lambda) = \lambda \log \mathrm{P(w|k)} + (1-\lambda)\log\frac{P(w|k)}{p(w)} \qquad (1)$$

where $P(w|k)$ is the probability of term *w* assigned to topic *k* and $P(w)$ is the marginal probability of term *w* in the corpus. The first component of the equation, $P(w|k)$, prioritizes terms with high frequency in a topic. However, it does not consider how unique term *w* is to topic *k*, which can be captured by $\frac{P(w|k)}{P(w)}$, a quantity that Taddy (2011) called *lift*. We set λ to 0.5 to take both components into consideration; λ determines the weight given to the probability of term *w* under topic *k* relative to its lift.

Topic Specificity. One issue with topic models is that it is not clear how topics vary in their association with immigrant corpora as compared with the entire corpus. We used Equation 2 to compute the specificity of topic *k* to the immigrant corpus:

$$\text{Specificity(k)} = \sum_{i=1}^{n}\left(\frac{\gamma(w_i|k)}{\sum_{i=1}^{n}\gamma(w_i|k)} * \frac{p(w_i|immigrant\ corpus)}{p(w_i|general\ corpus)}\right) \qquad (2)$$

[1] We used R lds library (Chang, 2012) to train the LDA model for multiple numbers of topics (from 10 to 20) using 1000 iterations. The hyper-parameters alpha and beta were set to 0.01 to encourage the model to assign topics to documents such that each document is composed of few topics and to learn topics that produce a few words with high probability.

where n is number of words assigned to topic k, $\frac{\gamma(w_i|k)}{\sum_{i=1}^{n}\gamma(w_i|k)}$ is the normalized relevance of word w_i to topic k, and $\frac{p(w_i|immigrant\ corpus)}{p(w_i|general\ corpus)}$ is the ratio of the frequency of word w in the immigrant corpus to its frequency in the source corpus. Specificity can range from 0 to near infinity. A specificity of 1 means that, on average, the words characterizing the topic have the same frequency in both the immigrant corpus and the source corpus. Larger topic specificity suggests stronger association to immigrant corpus.

Topic valence and concreteness. Topic valence and concreteness can be computed by averaging valence and concreteness ratings of every words that were assigned to each topic by LDA.

Distributions of topics over ethnic groups. To produce the strength of association between topics and immigrant groups, we computed the document-normalized probability distribution of words in ethnic corpora over the 15 topics, letting loading of an ethnic group on topic t be

$$l_t = \frac{\sum_{d\in D} P_{dt}}{\sum_{t\in T}\sum_{d\in D} P_{dt}}\ , \qquad (3)$$

where d is a document from an ethnic corpus D; t is one of the 15 immigrant topics. P_{dt} is proportion of words in document d that are assigned to topic t.

3.3 Results

Immigrant Sentiment and Concreteness. To understand the underlying mechanism driving sentiment towards immigrants, we computed the valence and concreteness for the language associated with each of the 56 immigrant groups (Fig 3.1). The most positively viewed ethnic group was African Americans and the least positive was Iraqi, with these two immigrant groups also showing high and low concreteness, respectively. This was reflected across ethnic groups as a whole, with language associated with more positively viewed ethnic groups being reliably more concrete ($r(55) = 0.77$, $p < 0.001$, 95%CI = 0.62 – 0.85). Overall, the pattern resembles Parrillo and Donoghue's (2005) finding among American college students that the various European groups continue to hold the highest degree of positive sentiment for immigrants, with a variety of Asian groups somewhere middle in the rankings, and groups from the Middle East continuing to rank near the bottom.

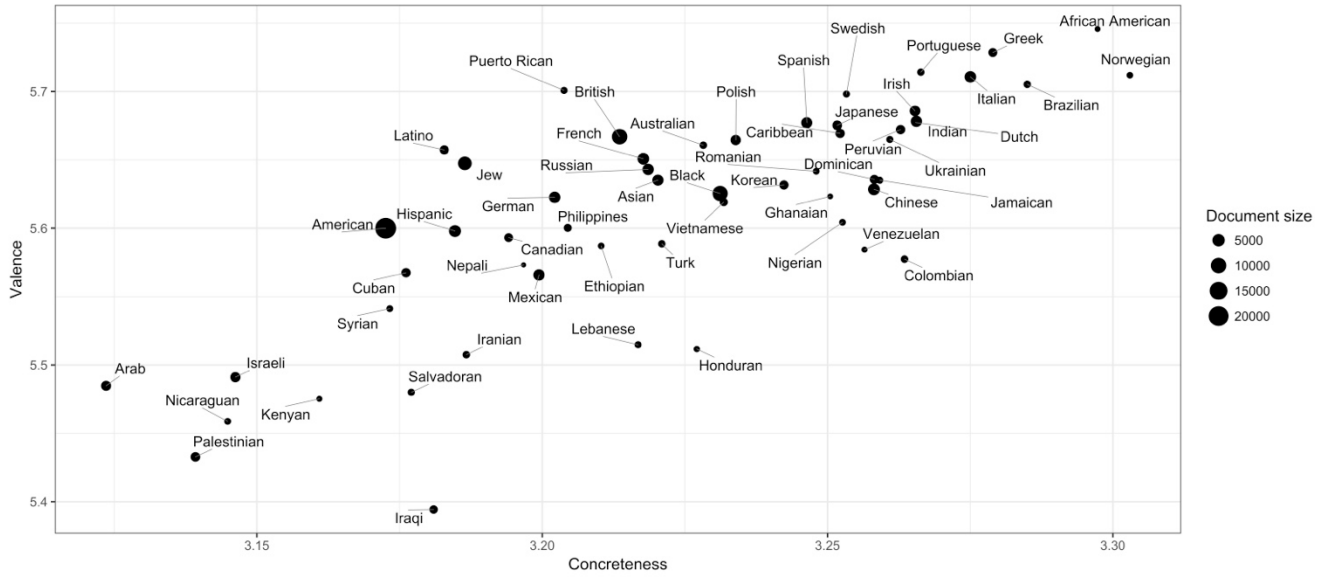


Figure 3.1. Relationship between valence and concreteness of ethnic corpora. The size of dots represents the size of each ethnic corpus.

To test the causal relationship between social distance and heightened sentiment, we divided the immigrant corpus into two time frames. Here we show the results for *T1* (from 1987 to 1994) and *T2* (from 2000 to 2007). Regression analysis was used to test if concreteness at T1significantly predicts change in valence between *T2* and *T1*:

$$Valence_{t2} \sim \alpha + \beta_{concreteness} * concrteness_{t1} + \beta_{valence} * Valence_{t1}$$

The results shows that the two predictors explains 69.3% of the variance ($R^2 = 0.69$, F(2,53) = 59.9, $p < 0.001$). Both concreteness and valence at *T1* significantly predict valence at *T2* ($\beta_{concreteness} = 0.38$, $p < 0.001$; $\beta_{valence} = 0.64$, $p < 0.001$). However, valence at *T1* was not a significant predictor of concreteness at *T2* in a corresponding regression including valence and concreteness at T1 ($\beta_{concreteness} = 0.70$, $p < 0.001$; $\beta_{sentiment} = 0.03$, $p > 0.10$).

These results hold for different values of *T1* and *T2*. Figure 2.A shows to what extent concreteness and valence at *T1* predict concreteness at *T2* when number of years included in *T1* and *T2* increases from 3 to 10. Both valence and concreteness are good predictors when at least four years are included in the time interval. In contrast, regardless of number of years included in time intervals, valence at *T1* is never a good predictor of concreteness at *T2* when valence at *T1* is controlled.

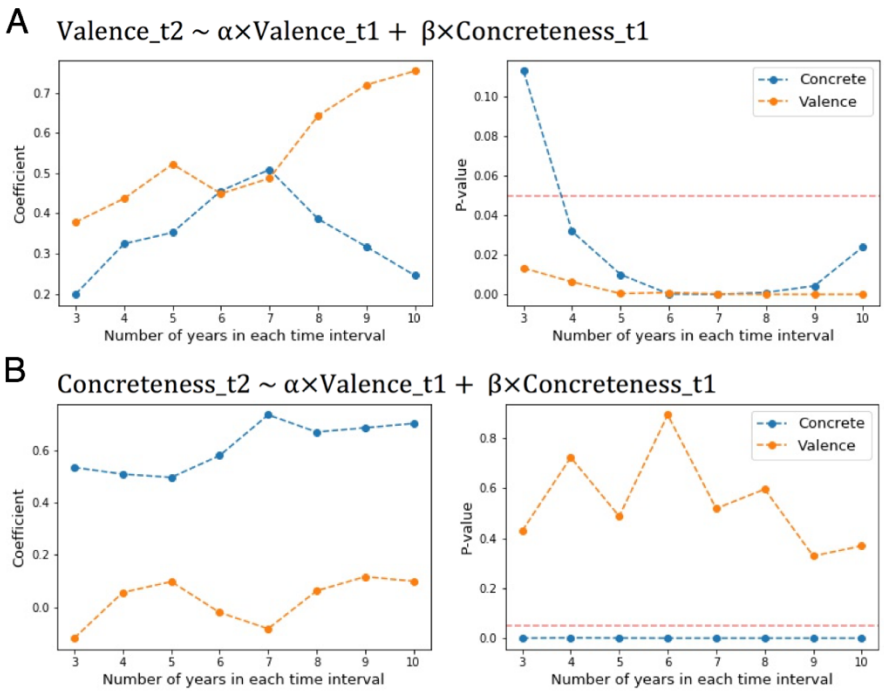


Figure 3.2. Sensitivity analysis: change of regression coefficients and p-values when number of years included at the beginning and the end varies from 3 to 10. A). Model that regresses valence of ethnic corpora at t2 on valence and concreteness at t1. B) Model that regress concreteness of ethnic corpora at t2 on valence and concreteness at t1.

Three additional hypotheses can be ruled out. First, at individual word level, we found a very weak positive correlation between valence and concreteness across the 13,384 English words in the Warriner et al. (2013) affective norms (Pearson's $r(13383) = 0.10$, $p<0.001$, 95% CI = 0.08 - 0.11). This is consistent with previous findings that concrete words are less neutral and more emotionally valenced (Vigliocco et al., 2013; Kousta, Vigliocco, Vinson, Andrews & Del Campo, 2011). At article level, the correlation between concreteness and valence across all immigrant news articles ($N = 43530$) is only 0.26 ($p < 0.001$, 95%CI = 0.25-0.27). Therefore, the large correlation we find across ethnic groups is unlikely to be an artefact of linguistic properties of English language. Second, if contact in appropriate situations reduces intergroup prejudice, frequency of exposure to outgroup information could potentially achieve a similar effect. However, we find no significant correlation between valence and media exposure, operationalized as number of articles that mentioned the target group (r (55) = 0.11, p =0.42, 95% CI = -0.16 – 0.36). Finally, it is possible that emphasis of immigrant identity leads to more negative attitudes towards ethnic groups. However, immigrant status, operationalized as the ratio between number of articles that mention an ethnic group as immigrants and all articles mentioning that ethnic group, is also not correlated with valence ($r(55) = 0.05$, $p = 0.72$, 95%CI

= -0.21 – 0.30). In other words, neither frequency of mentions nor immigrant status were sufficient to explain immigrant sentiment.

Table 3.1 Key words for each immigrant topic

Index	Topic	Key words
1	Police &crime	Police, officer, say, arrest, charge, prosecutor, drug, kill, gang, crime
2	Terrorism	Muslim, terrorist, bomb, attack, terrorism, intelligence, Islamic, FBI, mosque
3	Legal	Immigration, law, court, alien, judge, legal, justice, case, federal, lawyer
4	Politics	Republican, bush, democrat, bill, president, vote, senate, senator, campaign
5	Geopolitics	Israel, soviet, Socialist, Russian, France, Germany, Europe, Jew, Palestinian
6	Refugee	Refugee, Cuban, asylum, Cuba, Haitian, unite, Miami, boat, Castro, Haiti
7	Illegal worker	Worker, border, Mexico, company, Mexican, labor, job, wage, work, pay
8	Census	Hispanic, population, percent, Asian, black, census, Chinese, Korean, immigrant
9	Neighborhood	City, build, house, neighborhood, say, county, resident, island, apartment, rent
10	Book	Write, book, one, life, American, like, America, world, think, history
11	Religion	Church, catholic, Irish, bishop, priest, Jewish, religious, parish, pope, cardinal
12	Education	School student, child, teacher, education, parent, program, health, care
13	Restaurant	Restaurant, say, like, one, day, get, come, family, room, home
14	Music & Movie	Theater, film, music, movie, play, art, direct, musical, dance, song
15	Museum	Museum, Sunday, tour, street, information, tomorrow, admission, exhibition

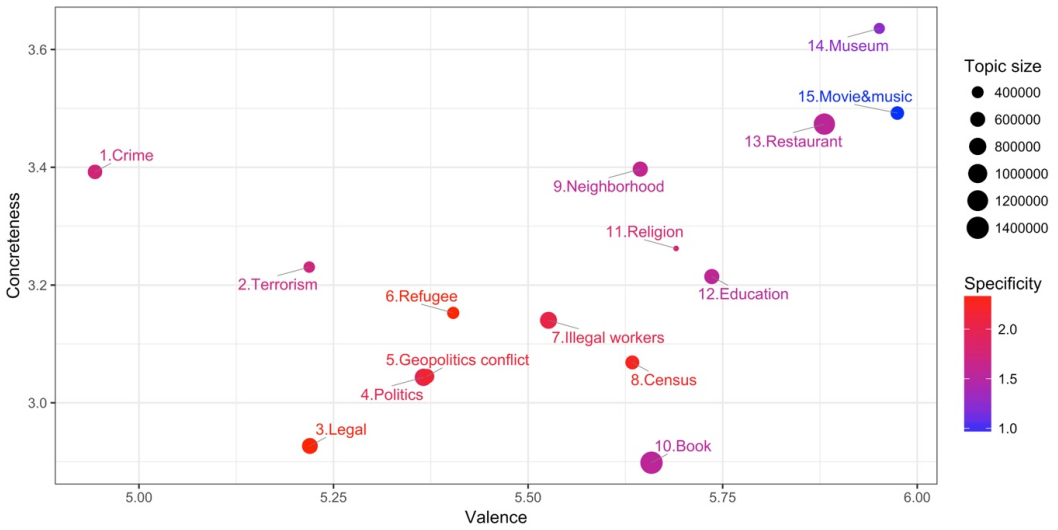


Figure 3.3. Concreteness and valence of the 15 immigrant topics identified using LDA. The dot size corresponds to the number of words assigned to that topic. The dot color represents topic specificity, with higher values indicating greater likelihood that a topic is used to refer to immigrants.

Immigrant Topics. We applied LDA to identify topics associated with the sentiment of ethnic groups. Table 3.1 shows the 10 most relevant words in each topic (see Equation 1 in methods section for definition of topic-word relevancy). Key words from the same topic are highly similar with each other and are clearly distinguishable from the words of other topics. We labelled topics by summarizing their keywords and present these labels in the *Topic* column. The results indicate a wide array of topics surrounding immigrants, with crime, terrorism and geopolitical conflict among the most negative topics while museum, movie and restaurant are among the most favourable. These topics reflect many of the issues commonly associated with both the pros and cons of immigrants (Cuddy, Fiske, Demoulin, & Leyens, 2000; Eckes, 2002; Cuddy et al., 2007; Peabody, 1985; Poppe, 2001; Alexander, Brewer and Herrmann, 1999; Borjas, 1990; Carens, 2013; Skeldon, 2014).

Next, we analysed linguistic features of these topics: valence, concreteness and topic specificity (Figure 3.3). We found no significant correlation between topic valence and concreteness ($r = 0.51$, $p = 0.052$). To capture the differences among topics in terms of their association strength with immigrants, we analysed topic specificity (see Equation 2), a measure of relative correspondence of each topic with the immigrant corpus as compared with the entire corpus. Larger topic specificity suggests words of a topic are more likely to appear in immigrant corpora than elsewhere in the NYT corpus. We found topic specificity is negatively correlated with concreteness ($r = -0.63$, $p = 0.012$) and valence ($r = -0.67$, $p < 0.01$). In other words, stronger association with immigrant leads topics to be more abstract and negative.

To understand what topics define media representation of each immigrant group, we computed the document-normalized probability distribution of words in ethnic corpora over the 15 topics (see Equation 3 in method section). Figure 3.4 presents the distribution of topics over ethnic groups. Groups ranked lower on mean sentiment are associated almost exclusively in a set of negative topics: Iraqis, Pakistanis, Syrians, and Lebanese are represented mostly in contexts of either terrorism or geopolitical conflicts; Nicaraguans, Vietnamese, and Venezuelans are closely associated with refugee topic while Mexican is associated most strongly with illegal workers. In contrast, groups ranked high in sentiment are closely associated with positive, and less immigrant-specific topics—such as restaurants, museums, and movies and music—and rarely represented with negative topics.

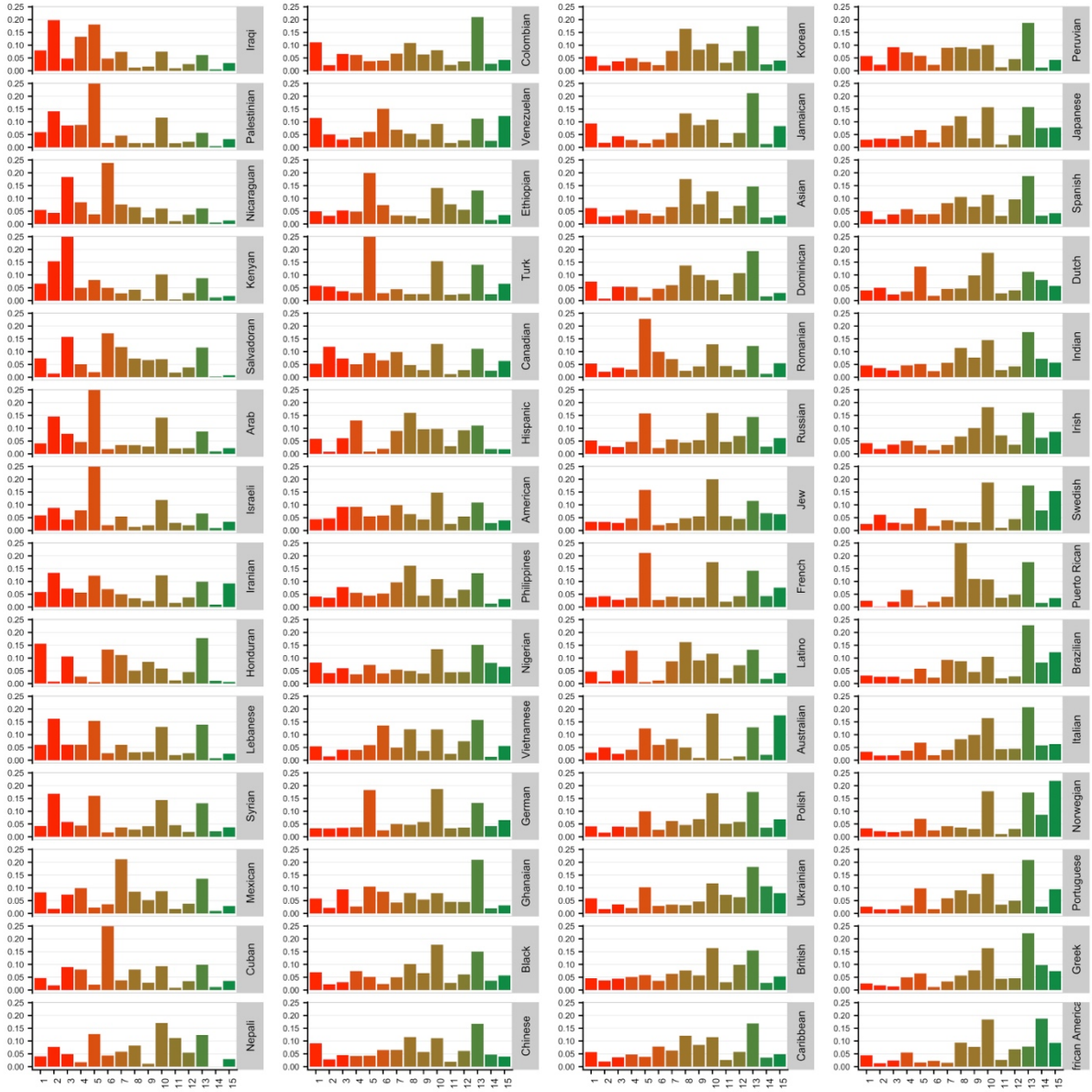


Figure 3.4. Distributions of topics over ethnic group ranked by valence. The x-axis shows the index of topic numbers identified in Table 4.1. The y-axis shows the normalized weighting of each topic on each immigrant group. Topics are arranged by valence, with the lowest (in red) on the left and the highest (in green) on the right; immigrant groups are also ranked by their overall valence, with the most negative group on the top left corner and the most positive on the bottom right.

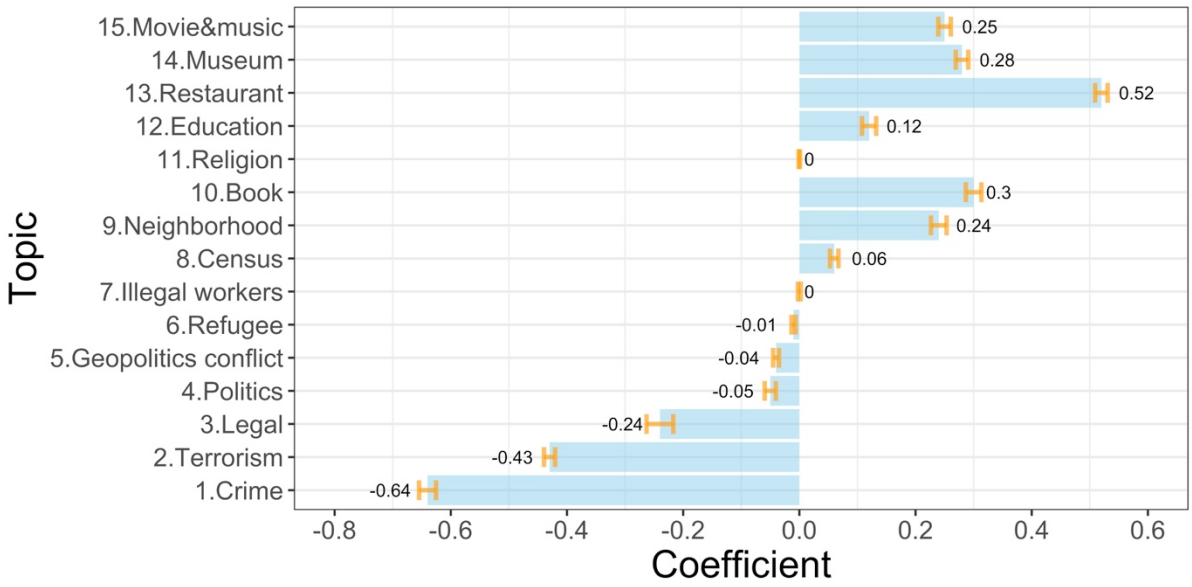


Figure 3.5. Regression coefficients of immigrant topics in an averaged linear regression model that predicts valence of immigrant groups using the distribution of immigrant language over the 15 topics. The error bar represents 95% of confidence interval.

To quantify what topics are driving the sentiment of immigrants, we regress valence of immigrant groups on the distribution of immigrant language over 15 topics as input. Since this model contains 15 independent variables and only 56 data points, we used elastic net regularization, which is a combination of Lasso regression and ridge regression. These two perform simple linear least squares, but penalize the coefficients of the inputs x_i based on their size. The penalty forces some regression coefficients to zero. To assess the fit of the model, we performed the cross-validation exercise. We divided our data set into 10 equal groups, trained our model on a random sample of 7 groups, then predicted immigrant sentiment in the remaining 3. This cross-validation exercise was repeated 1,000 times to calculate the averaged adjusted R^2 for the out-of-sample predictions and averaged regression coefficients. We found that the topic profiles of immigrant groups are able to capture 78% of the variance of their sentiment. Overall, the negative topics have stronger impact than the positive topics (Figure 5). Crime, Terrorism, Legal, are the three major negative topics that pull down the perceived sentiment of immigrants. Restaurant is the most effective topic at boosting sentiment. Surprisingly, Politics, Geopolitical Conflict, Refugee, Illegal Workers, and Religion fail to emerge as important predictors of future immigrant sentiment.

3.5 Discussion

Among studies on prejudice and stereotype, mass opinion on immigrant groups is arguably the least studied quantitatively, perhaps due to its extensive cost. The present study

makes two contribution to this area of research. First, we found perceived social distance towards an immigrant group, operationalised by abstractness of the language, predicts future change in sentiment. Many studies of outgroup negativity have routinely found that prejudice is associated with perceived social distance (Gilbert, 1998). Our research extends this framework by demonstrating its role in predicting the sentiment of 56 ethnic groups, and by showing change in social distance precedes subsequent sentiment change of outgroups. Going beyond the two dimensions of language, we further identified topics associated with each immigrant group and how these topics were related to changing sentiment.

We also show that "immigrant" as a concept is representative not of a categorical variable, but rather as a continuous measure, to which essentially all North American's belong. Classic theories on outgroup negativity have been focused on a dichotomy of ingroup vs outgroup. Therefore, they often fail to explain differences between outgroups. By offering a continuous conceptualisation and measurement of relationship among outgroups, our study is able to explain which outgroups are more psychologically distant than others and explain how such differences lead to their future change in perceived sentiment.

The fact that our finding on social distance is largely consistent with Parrillo and Donoghue's survey (Parrillo & Donoghue, 2005) suggests that our corpus approach captures meaningful patterns despite possible limitations (e.g. representativeness). For example, being the 2nd largest news distributor in the U.S, and headquartered in a metropolitan city, the NYT is well-positioned to offer wide coverage on immigration issues and to influence its readers' attitudes towards outgroups. We acknowledge that immigrant topics may differ across media that target audiences in different parts of the world. However, we anticipate the relationship we found between social distance and perceived sentiment are generalizable to wider contexts, such as languages produced during daily conversations or on social networks like twitter. Unlike the NYT, these channels may be less restricted to use formal and politically correct language, leading to socially distant outgroups associated with even more negative language.

Overall, the utility of corpora approach outweighs its limitations. It has the advantage of (a) ecological validity through observation of psychological distance and sentiment in texts produced on a variety of topics, (b) tracking relationships between psychological distance and sentiment in language referencing a large variety of ethnic groups, (c) allowing researchers to study perceptions on immigrants outside the laboratory and avoid problems such as receiving socially desirable answers, and (d) providing information over time (20 years in the present case) allowing researchers to test causal relationships.

Chapter 4 : Words to Cultural Change

A cultural history of risk

Understanding how societies have conceptualized risk throughout history may help to predict the public's response to current and future threats and dangers. However, reconstructing what has been perceived as a risk over different historical periods is a thorny task. Previous investigations have commonly taken a qualitative approach. Complementing this approach, we propose that the historical dynamics of the conceptualization of risk can also be studied by analyzing the language used to construct and single out risks. Drawing on two large corpora, the Google Books Ngram Corpus and The New York Times Annotated Corpus (NYT Corpus), we take both a telescopic view and a more microscopic view. The former permits us to survey interpretations of risk over the past last two centuries; the latter focuses on the recent past from 1987 to 2007. Our analyses show that the construct of *risk*, unlike its synonyms *danger* and *hazard*, has undergone enormous conceptual change over time. Over the past two centuries, the word *risk* has been used increasingly frequently; it has appeared in a wider range of contexts; and the sentiment carried has become increasingly negative. In terms of risk topics, concerns over violent death (war) have, for the most part, been replaced by references to the risks of modernity, such as chronic disease and threats to the economy. These results offer quantitative insights into the cultural history and transformation of a multidimensional construct. They may further inform expectations about the dynamics of the future public discourse on risk in unsettled times.

4.1 Introduction

Humans have always been exposed to risks. Yet the nature of these risks has changed profoundly over the course of human biological and cultural evolution. Whereas the dominant risks were once starvation, infections, and violent conflict (Harari 2015), many of today's risks are associated with lifestyle choices (e.g., obesity, cardiovascular disease, cancer). Although modern institutions such as hospitals, police and fire services, and international treaties now buffer people in industrialized nations from the worst consequences of risks, the "consequences of modernity" (Giddens 1990) include new risks, such as nuclear weapons, global pandemics, deadly hospital bugs, fundamentalist terrorism, cyberattacks, and climate change. Despite

reductions in the rates of violent conflict, poverty, and starvation (Pinker 2011) and a doubling of life expectancy over the past two centuries (Oeppen and Vaupel 2002), many people appear to feel that the world is more rife with dangers than ever (see Pinker and Mack 2014). Indeed, the historian Bourke (2005) has argued that "fear is the most pervasive emotion of modern society." Relatedly, life in today's "risk society" (Beck 1992) seems to be characterized by rising vigilance to a growing variety of risks and insecurities (e.g., the precautionary principle; Sunstein 2005).

How does society identify risks? Cultural anthropologists and sociologists have emphasized that risks are not a natural kind but are socially constructed, based on norms, moral considerations, and structures of social organization (Douglas 1992). What qualifies as a risk is therefore subject to dynamic social change. For instance, today's religiously motivated terrorism is a striking example of how an "old" risk transforms into a new phenomenon and forcefully reappears on the collective radar. Bourke (2005) has documented a history of fears, from the Victorians' dread of being buried alive to the more recent fear of nuclear annihilation. These fears are preserved in cultural artifacts such as books and newspaper articles—records that provide insights into how risks are collectively identified and perceived. Taking a historical perspective on these artifacts reveals how and why society's attitudes to risk have changed and may indicate how they will change again in the future. Our goal is to take a large-scale quantitative approach to the recent historical trajectory of the word *risk* with the aim of understanding the changing nature of its social construction.

Before we turn to our research questions, let us clarify that the term *risk* is often used to mean different things. In the risk management and actuarial literature, for instance, it describes a loss of a certain magnitude (e.g., injury, mortality) weighted by the probability of its occurrence (Short Jr 1984, Rayner and Cantor 1987). By this actuarial measure, driving is riskier than flying because it is associated with a greater risk of injury per mile travelled. In the economic discourse, risk commonly refers to the variance in possible (positive or negative) returns. For instance, an investment option with higher return variance is deemed as riskier than an option with lower variance but the same expected mean return (Markowitz 1952, Pratt 1964). Research in psychology, sociology, and anthropology has consistently demonstrated that these actuarial and economic definitions are too narrow to capture people's understanding of risk. Lay perceptions are multidimensional, encompassing higher order factors such as *dread* and *equitable exposure* (Slovic 1987, Bhatia 2019). Dread risks (as opposed to "chronic risks") are defined by a perceived lack of control and potential large-scale loss of life, making flying a greater perceived risk than driving (e.g., Gaissmaier and Gigerenzer 2012). Greater dread, in

turn, is associated with greater perceived risk and a greater desire for regulation to reduce the risk (Slovic et al. 1985, Slovic 1987, Sunstein 2005). All these meanings and others are part of the public discourse and are included in the text corpora that we analyze. In other words, our focus is not on one definition at the expense of another, but rather endorses the rich and inclusive semantic history of *risk* in the natural language.

4.2 Guiding research questions

Our goal in this article is to track change in the public discourse on risk over historical time by addressing four guiding questions. First, we examine how the frequency of the word *risk* has changed over historical time. Word frequency has been used to capture patterns of usage associated with changes in cultural importance (Twenge et al. 2012, Greenfield 2013, Uz 2014). Here, it allows us to evaluate the idea that the construct of risk is playing an ever-increasing role in the public discourse. Second, we investigate how the sentiments of the words co-occurring with *risk* have changed. This sentiment analysis allows us to evaluate the hypothesis that risk is becoming a more negative construct and that societies and policy makers should perhaps invest more in risk reduction and prevention (the precautionary principle; Sunstein 2005). Third, we ask how the meaning of *risk* has changed by examining change in the semantic relationship between it and other words. The meaning of a word can be reliably inferred from the contexts in which it has been used (Firth 1957). For example, analysis of the linguistic contexts of *broadcast* shows that 150 years ago it referred to the spreading of seed, while it is now used to mean the spreading of information (Li et al. 2019). We examine the text corpora for indications that *risk* is more subject to semantic change than close semantic associates such as *danger* and *hazard*. Fourth, we decompose the construct of *risk* into the specific topics with which it has been associated and track those topics over historical time. Our purpose here is to identify the most prominent risk topics over time and to consider how they have changed in relation to world events.

We investigated these questions by analyzing latent semantic patterns in natural language. Tracing the historical meanings of words requires a corpus of texts published over a sufficiently long time period. The Google Books Ngram Corpus (Lin et al. 2012) is one of the few corpora that meet this requirement. Drawing on over 100 sources (e.g., libraries and publishers), it contains over 8 million books published from 1600 to 2008, or 6% of all books ever published. The corpus thus offers a *telescopic view* over a large time period. The corpus has been used to detect large-scale changes in language, which in turn correlate with social and demographic changes (Michel et al. 2011, Hills et al. 2012, Hills and Adelman 2015, Hills et

al. 2015). Any corpus, however, has its limitations. The Google Books Ngram Corpus offers limited contextual information due to a narrow window size (5-grams, or a contiguous sequence of five words); moreover, there has been a surge in the proportion of academic articles in the corpus (Pechenick et al. 2015). We therefore also examined *The New York Times Annotated Corpus* (NYT corpus; Sandhaus 2008) to allow cross-validation of the results. This corpus contains all (1.8 million) articles published in the *New York Times* from 1987 to 2007, and offers a more *microscopic view* on the risks of modern life as reported in the most widely read U.S. newspaper. Let us emphasize that because our analysis draws on English texts only, the present results are limited to English-speaking cultures. In addition, the two corpora can of course provide only a limited window onto the public discourse on risk. Nevertheless, the Google Books Ngram Corpus, in particular, has the advantage of covering a relatively long time period, going beyond short-term analyses of, for instance, media coverage of risk and mortality (see the references in Young et al. 2008).

4.3 Materials & Methods

Google Books Ngram Corpus. The Google Books Ngram Corpus consists of *n*-grams: contiguous sequences of *n* items from a given text (*n* ranges from 1–5). We used the 5-grams of all English words in our analysis; each data entry therefore displays the number of times a 5-gram appears in the corpus during a specific year. We retrieved all 5-grams starting or ending with the word *risk*. As is standard data-cleaning procedure in many natural language processing tasks, we removed stop words, punctuation, digits, and words containing fewer than three characters before using the WordNet-based NLTK lemmatizer (Bird, Loper, & Klein, 2009) to lemmatize each noun to its singular form and each verb to its present tense. Next, we aggregated all 5-grams by year so that all words appearing in the same year were treated as one document. Aggregating topics by years encourages the topic model to identify the underlying patterns that best explain differences among risk structures over years.

The New York Times Annotated Corpus. The NYT Corpus contains uncontracted news articles. We constructed a risk corpus by selecting articles that mentioned the word *risk* or *risks* more than twice. Next, we pre-processed the corpus in the same way as we did the Google Ngram data except for aggregating articles by year: Each news article was treated as one document.

Analysis of Frequency, Contextual Sentiment, and Semantic Drift. Analyses of frequency, contextual sentiment, and semantic drift (Fig. 4.1 and Fig. 4.2) were conducted using the Macroscope (Li, Engelthaler, Siew & Hills, in press), an interactive linguistic tool

that facilitates analysis of historical sentiment and semantic change. It was built using the historical data made publicly available by the Google Books Ngram Corpus. Refer to the SI for details of the procedure.

Topic Modelling. We studied historical change in the meaning of the word *risk* by extracting risk topics from two large corpora: the Google Books Ngram Corpus (Lin et al., 2012) and the New York Times Annotated Corpus (Sandhaus, 2008). The topic model we used was latent Dirichlet allocation (LDA; Blei, D.M., Ng, A. Y. & Jordan, M. I., 2003), a bag-of-words algorithm that identifies a set of topics that best describe/re-generate the corpus. We took two main steps in analyzing the data. First, we identified the structure of risk meanings by applying the topic model to the risk corpus. This step allowed us to understand the key events associated with risk. Next, we applied trend analysis to understand how the risk topics identified in the first step changed over time. See the SI for details on the implementation of LDA.

Interpreting Topics. To make sense of the meanings of the risk topics, we used Equation (1) to identify the words most relevant to each topic. The relevance of term *w* to topic *k* given a weight parameter λ was defined as:

$$\gamma(w, k|\lambda) = \lambda \log\left(P(w|k) + (1-\lambda)\log\left(\frac{P(w|k)}{P(w)}\right)\right), \qquad (1)$$

where $P(w|k)$ is the probability of term *w* assigned to topic *k* and $P(w)$ is the marginal probability of term *w* in the corpus. The first component of the equation, $P(w|k)$, prioritizes terms with high frequency in a topic. However, it does not consider how unique term *w* is to topic *k*, which can be captured by $\frac{P(w|k)}{P(w)}$, a quantity that Taddy (2011) called *lift*. We set λ to 0.5 to take both components into consideration; λ determines the weight given to the probability of term *w* under topic *k* relative to its lift.

One issue with topic models is that it is not clear which topics capture structures specific to the risk corpus and which topics capture general features of the source corpus. To find out, we used Equation (2) to compute the specificity of topic *k* to the risk corpus:

$$\text{Specificity(k)} = \sum_{i=1}^{n}\left(\frac{\gamma(w_i|k)}{\sum_{i=1}^{n}\gamma(w_i|k)} * \frac{p(w_i|risk\ corpus)}{p(w_i|general\ corpus)}\right), \qquad (2)$$

where $\frac{\gamma(w_i|k)}{\sum_{i=1}^{n}\gamma(w_i|k)}$ is the normalized relevance of word *w* to topic *k*, and $\frac{p(w_i|risk\ corpus)}{p(w_i|general\ corpus)}$ is the ratio of the frequency of word *w* in the risk corpus to its frequency in the source corpus. Specificity can range from 0 to almost infinity. A specificity of 1 means that, on average, the

words characterizing the topic have the same frequency in both the risk corpus and the source corpus, suggesting that the topic reflects the underlying pattern of the source corpus, not risk. An example of a nonspecific topic is one that generates words necessary to construct every document, such as articles and pronouns. The absolute value of topic specificity is heavily influenced by the data format: NYT articles are more likely than 5-grams to contain non-risk-specific words (noise) and therefore have smaller values of $\frac{p(w_i|risk\ corpus)}{p(w_i|general\ corpus)}$. Topic specificity is not comparable across corpora; instead, it should be used to compare topics from a same corpus.

Tracking Trends in Topics. To analyze trends in topics over time, we used the output from the LDA model on the Google Books Ngram Corpus to calculate the contribution of each topic *k* in each year using Equation (3). For each document (i.e., all 5-grams in a specific year), the equation controls for document length by dividing the number of words generated by each topic by the total number of words in the document. Thus, the yearly topic contribution estimate, $p_d(k)$, is defined as:

$$p_d(k) = \frac{|\{w \in d: topic(w) = \mathrm{k}\}|}{|d\ |}, \qquad (3)$$

where *k* is a topic and *w* is a word in a document *d*. The numerator is the number of words in document *d* that are generated by topic *k*; the denominator is the total number of words in document *d*.

4.4 Results

How Has the Frequency of *Risk* Changed Over Time?

We first investigated change in the frequency of the word *risk* over time, starting with the Google Books Ngram Corpus. As Figure 1A shows, use of the word *risk* has increased dramatically since about 1970, with an approximately fourfold increase in usage since the 1950s. We checked this trend in English against other languages and found similar increases in French, German, Italian, and Spanish (Figure 1B). In addition, we observed a similar proliferation of *risk* in the Corpus of Historical American English (COHA; Davies 2008). As COHA is balanced by genre and subgenre across decades,[2] these findings suggest that *risk* proliferation is not an artifact of increasing numbers of scientific journals being included in the Google Books Ngram Corpus (Figure 1A). There is, however, no sign that the public discourse has turned darker in general, as close semantic relatives signifying undesirable states such as

[2] For example, fiction accounts for 48–55% of the total in each decade (1810s–2000s); subgenres such as prose, poetry, and drama are likewise balanced. This balance across genres and subgenres means that researchers can be reasonably certain that patterns in the data do not merely reflect artefacts of a changing genre balance.

fear, *danger*, and *hazard* are not being used more frequently. On the contrary, the use of *fear and danger* has declined steadily over the past two centuries, while the use of *hazard* has remained relatively stable at a low frequency. These results are consistent with the idea that *risk*, more than other terms, has become central to the public discourse (Beck 1992, Bourke 2005).

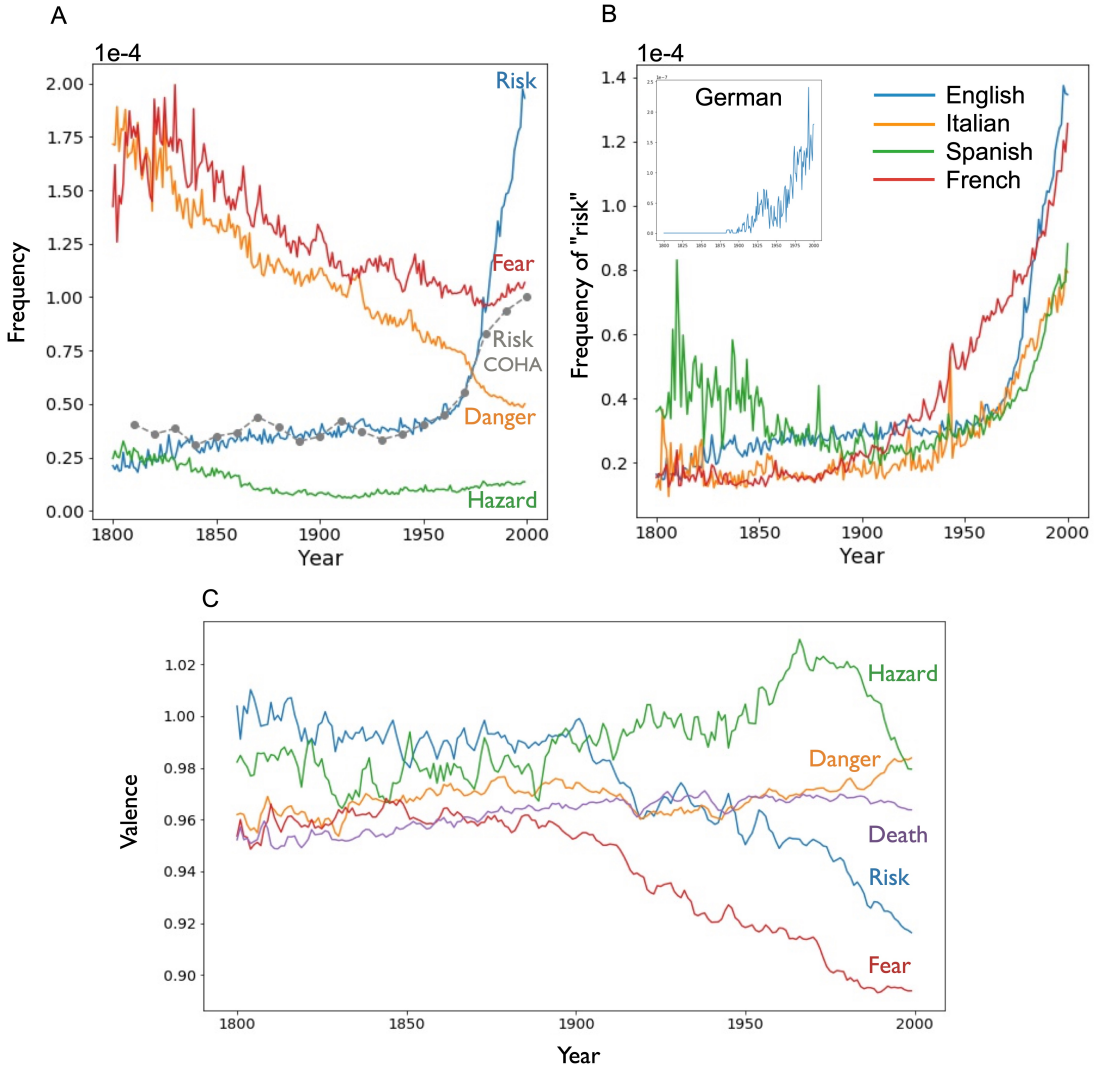


Figure 4.5. Historical change in the frequency and sentiment of the word *risk* and its close semantic neighbors in the Google Books Ngram Corpus. (A) Frequency of *risk*, *fear*, *danger*, and *hazard* in the Google Books Ngram Corpus and frequency of *risk* in the Corpus of Historical American English (COHA). (B) Frequency of *risk* in five languages—English, Italian, Spanish, French, and German—in the Google Books Ngram Corpus (C). Change in the sentiment of words co-occurring with *risk, fear, danger, hazard*, and *death*. Sentiment was adjusted to mean score of all words, such that valences > 1 indicate a more positive context than average. The word *death* is included to provide a sentiment benchmark, as its meaning and sentiment have remained stable over history.

How Have the Sentiments Associated with *Risk* Changed?

Next, we examined whether the sentiments associated with *risk* have changed over time. For example, is it possible—in line with a more economic interpretation of risk—that the use of the word *risk* is increasingly associated with an appreciation of the large potential rewards that make some risks worth taking (Hertwig and Pleskac 2014)? This is not the case, as the

results presented in Figure 1C show. Computing the weighted average valence of the words that co-occurred with *risk* over the past 200 years revealed that the sentiment associated with risk has become increasingly negative, showing a roughly monotonic decline from 1800 to 2000. To provide points of comparison, we also analyzed related concepts (*fear, danger, hazard*) as well as *death* as a benchmark. The sentiment analysis shows that *risk* has undergone a much larger change over time than these inherently undesirable concepts (with the exception of *fear*). In the early 1800s, the sentiment of words co-occurring with *risk* was more positive than that of any of the four comparison words; by the end of 20th century, it was more negative than that of *danger, hazard,* or *death* (Figure 1C). In other words, the word *risk* has become not only more prevalent but also more negative in meaning.

How Have the Semantic Relationship of Risk Changed?

Next we turn to an analysis of semantic drift, which likewise suggests that *risk* has experienced more change over historical time than its close semantic relatives. Specifically, Figure 2 visualizes the semantic associates of *risk*, *danger*, *hazard*, and *fear* in two-dimensional space relative to their *k* most similar words in 1800 and 2000 ($k = 7$ for each word). The pattern is clear: *risk*, *danger*, and *hazard* started as close semantic neighbors in 1800 and moved apart over time. By the year 2000, the underlying semantics of *risk* had grown more similar to those of *prevalence* and *prevention*, terms associated with the quantification, reduction, and avoidance of risk. *Danger* and *hazard*, in contrast, remained in the semantic area defined by words such as *harm, threat*, *adverse*, and *peril*. This finding suggests that the word *risk* has moved from representing the existence of threats to describing their examination, quantification, and prevention.

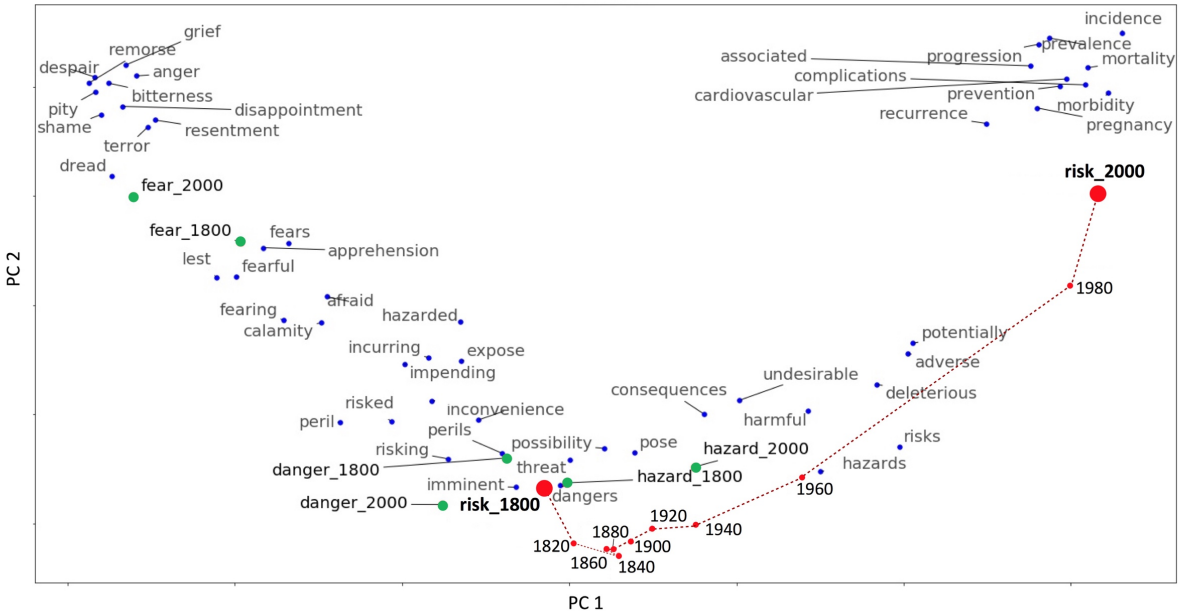


Figure 4.6. Semantic drift of *risk*, *hazard*, *danger*, and *fear* from 1800 to 2000 in the Google Books Ngram Corpus. The target words (*risk* as red dots; the other three as green dots) are shown in relation to their near associates (as blue dots) in the years 1800 and 2000. The words are presented in two-dimensional space based on their word embeddings. The axes represent the two dominant principal components. The words *risk*, *danger*, and *hazard* started as near neighbors in 1800 but moved apart over time.

How Have Risk Topics Changed Over Time?

The semantic drift analysis shows how *risk* has diverged from its semantic neighbors over the last two centuries, but it fails to capture the topical dimensionality of risk in this period. As noted by Blais and Weber (2006), risk is a multidimensional concept encompassing numerous topics. We therefore applied LDA to investigate the topics that have driven the proliferation of *risk* in the public discourse and its increasingly negative sentiment. We inferred topic meanings by inspecting their most relevant words (see Equation 1 in the Methods section), as summarized for each topic in Table 1. Applying the topic model to the Google Books Ngram Corpus identified six risk categories: **war** (topic 1, 2, 3), **nuclear** (topic 4), **health** (topic 5, 6, 7, 8, 9), **HIV/AIDS** (topic 10, 11), **risk society** (topic 12), **economy** (topic 13, 14), and a non-specific topic on risk analysis (topic 15).

Table 4.2. Most Relevant Words for Each Risk Topic, Ordered by Relevance as Defined in Equation 1

Index	Google Books Ngram Corpus	NYT Corpus
1	Life, imminent, battle, resolve	Military, war, Iraq, troop
2	Life, war, bureau, loss	China, Japan, country, foreign
3	War, uncertainty, loss, prepare	Environmental, plant, energy, gas

4	Nuclear, carcinogenic, patient, infant	Cancer, woman, study, breast
5	Heart, coronary, injury, bear	Drug, patient, doctor, hospital
6	Breast, cancer, osteoporosis, fetus	AIDS, virus, infect, vaccine
7	Stroke, cancer, disease, capital	Child, school, parent, student
8	Prostate, cancer, event, Alzheimer	Fund, stock, investor, market
9	Management, diabetes, cardiovascular, overweight	Food, fat, eat, diet
10	AIDS, nation, HIV, immunodeficiency	Insurance, bank, loan, insurer
11	HIV, deficit, assess, volume	Law, court, abortion, tobacco
12	Management, value, assessment, society	Airline, flight, shuttle, space
13	Confrontation, return, equilibrium, preference	Company, business, executive, industry
14	Rate, free, interest, return	Investigation, Enron, prison, police
15	Behavio[u]r, group, death, population	Think, people, way, thing
16		Republican, Clinton, Bush, Democrat
17		Game, player, sport, team
18		Day, car, hour, walk
19		City, build, York, new
20		Film, art, movie, theater

Note: As the topics are ordered by relevance, they are not aligned for the two corpora here. Figure 4.3 shows the topics aligned for the two corpora and by risk category.

Each topic represents a probability distribution over all words. In order to validate our interpretation of risk topics from the Google Books Ngram Corpus, we selected a collection of words (see the left column of Figure 3A) that characterize each of the risk categories identified above and examined how those words were distributed over topics (see the left panel of Figure 3A). The words were selected by referring to the list of most relevant words for each topic and screening out generic words such as *significant*, *total*, and *factor*. Topics from the same category are more likely to generate corresponding words but not others. This pattern, visualized as probability loadings on the diagonal of the word-topic probability heat map in Figure 3A, supports the interpretation of topic meanings in Table 1.

How replicable is this category structure? To find out, we also analyzed the NYT Corpus. Applying the same procedure to the NYT Corpus confirmed all risk categories inferred for the Google Books Ngram Corpus (visualized as probability loadings on the diagonal of the right panel of Figure 3A). We can therefore conclude that the meanings of risk derived in our analysis of the Google Books Ngram dataset are not corpus-specific results associated with a non-representative sample, but reflect general trends in the topicality of risk over both long and short time scales.

In order to ensure that the topics were risk-specific and did not just reflect the background features of the corpus, we next computed *topic specificity* (see Equation 2 in the Methods section) to quantify the relative correspondence of each topic with the risk corpus as

compared with the entire corpus (see Figure 3B). A topic specificity score around or below 1 means that the topic has a distribution of words similar to that seen in the entire corpus; the topic therefore represents the general features of the entire corpus. For the Google Books Ngram Corpus, we found the topic specificity of all risk topics to be above 1 (ranging from 50 to 650), suggesting that all topics were risk-relevant. In contrast, the specificity of NYT topics ranged from 0.7 to 2.5, with six topics being irrelevant to risk (the specificity scores of topics 15–20 were close to or less than 1). This notable difference in the topic specificity of the two corpora may be attributable to differences in data format: Recall that the Google Books Ngram data contain words that co-occurred with *risk* within a narrow window size, whereas the NYT data contain entire articles that mention the word *risk*. As such, NYT articles are more likely than Google Books Ngrams to contain words not specific to *risk*.

Nevertheless, both corpora rendered a similar set of high-specificity topics: nuclear, heart disease, cancer, diabetes, and HIV/AIDS. War-related topics had low specificity in the NYT Corpus. This result is not surprising because, as we show in the following analysis, war topics have gradually disassociated from *risk* since World War II, and the NYT Corpus only dates back to 1987. Beyond the risk topics identified for the Google Books Ngrams, we found only one additional topic in the NYT Corpus with specificity clearly above 1 (topic 9, featuring words such as *food, fat, eat,* and *diet*), and four additional NYT topics slightly above 1 (topics 11–14, which we interpreted as legal, flight, commercial, and fraud, respectively). Correspondingly, the key words associated with topics 11–14 showed low co-occurrence with *risk* in the Google Books Ngram Corpus throughout history. This comparison suggests that, overall, both corpora converged on a similar set of important risk categories.

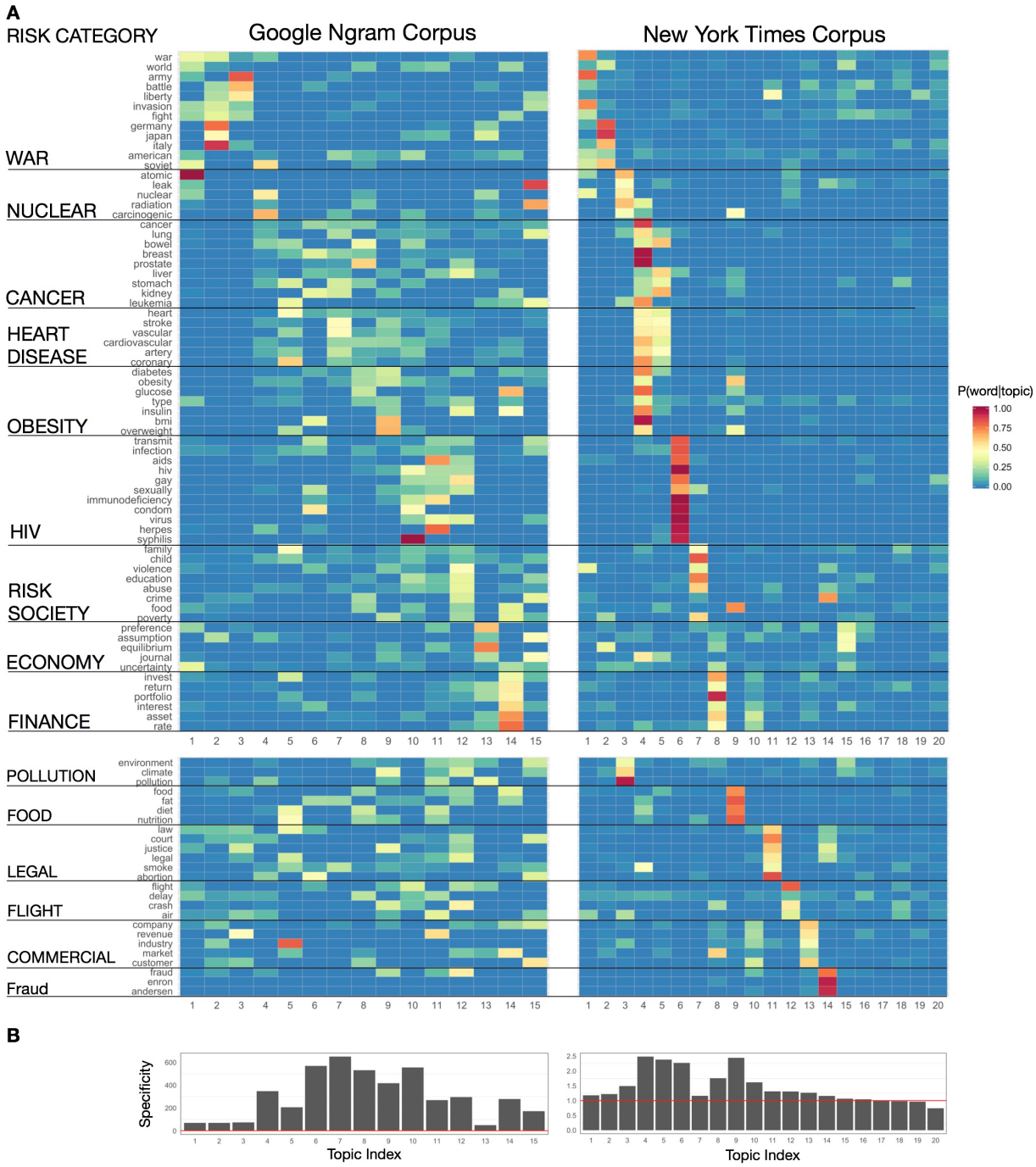


Figure 4.7. Visual quantification of risk topics. (A) Heatmap of the probability that word *w* was generated by topic *k* in models derived from the Google Books Ngram Corpus (left) and the NYT Corpus (right). Words on the y-axis were selected by referring to the list of most relevant words for each topic (relevance defined by Equation 1) and they were grouped by categories. (B) Topic specificity (as defined by Equation 2). The red horizontal line indicates topic specificity equal to 1. Topics with specificity above this reference line can be considered risk-specific and therefore capture one or more aspects of the meaning of risk. Topics with specificity below 1 can be considered generic words that are not informative with respect to risk meanings.

How Are Changes in Risk Categories Associated With Other Events and Developments?

One advantage of Google Books Ngram Corpus is that it allows us to investigate change in the meaning of risk across a period of over 150 years and to speculate on how those changes

relate to other historical events and developments. Specifically, we performed a trend analysis on the topic model derived from the Google Books Ngram Corpus over the years 1850 to 2008. As Figure 4 shows, the structure of the Google Books Ngram risk topics underwent major changes over this period. The three war-related topics emerge early in the distribution: Topic 1 (*life, imminent, battle, resolve*) dominated the risk structure in the second half of the 19th century, which witnessed several major wars (e.g., Crimean War, American Civil War). Topic 2 (*life, war, bureau, loss*) emerged and reached its peak during World Wars I and II. Topic 3 (*war, uncertainty, loss, prepare*) reached its peak during the Vietnam War. Topic 4 (*nuclear, carcinogenic, patient, infant*) peaked around 1985, capturing the risks associated with the proliferation of nuclear weapons during the Cold War (see the histogram in Figure 4) and the growing use of nuclear power in the 1970s and 1980s.

Chronic diseases such as heart disease and cancer are now the leading global risks for mortality (World Health Organization 2009). Topics reflecting this development (topics 5–9) started to emerge from the 1970s and remain the most prominent risk topics. Due to the large proportion of shared words associated with the different health conditions, topics 5, 6, 7, and 8 show considerable overlap, that is, they share words that describe cancer, heart and coronary issues, and other severe diseases. Topic 9, associated with obesity and diabetes, emerged after 2000. The data for topics 10 and 11 show that concerns over AIDS and HIV emerged within 2 years of the first AIDS diagnosis in the US in 1981 and soon reached a peak around 1995, when the reported annual mortality from HIV/AIDS peaked in the United States (CDC 1999, 2003, 2006, 2010). Potentially reflecting the fact that an HIV diagnosis no longer represents a death sentence, this risk topic decreased in prominence after 2000 (see the histogram of AIDS-related deaths in the US in Figure 4).

Finally, topic 12 (*management, value, assessment, society*) is about management of various social risks. It seems to relate to Beck's conceptualization of the *risk society*, being associated with words such as *Ulrich, Beck,* and *modernity*. Topics 13 and 14 relate to the economy, and emerged from the 1970s: topic 13 features words like *preference*, *assumption, equilibrium*, and *journal*, whereas topic 14 features words such as *return*, *portfolio,* and *interest*. Lastly, topic 15 (*behavior, group, death, population*) seems to be concerned with general risk analysis, without reference to any specific risk event.

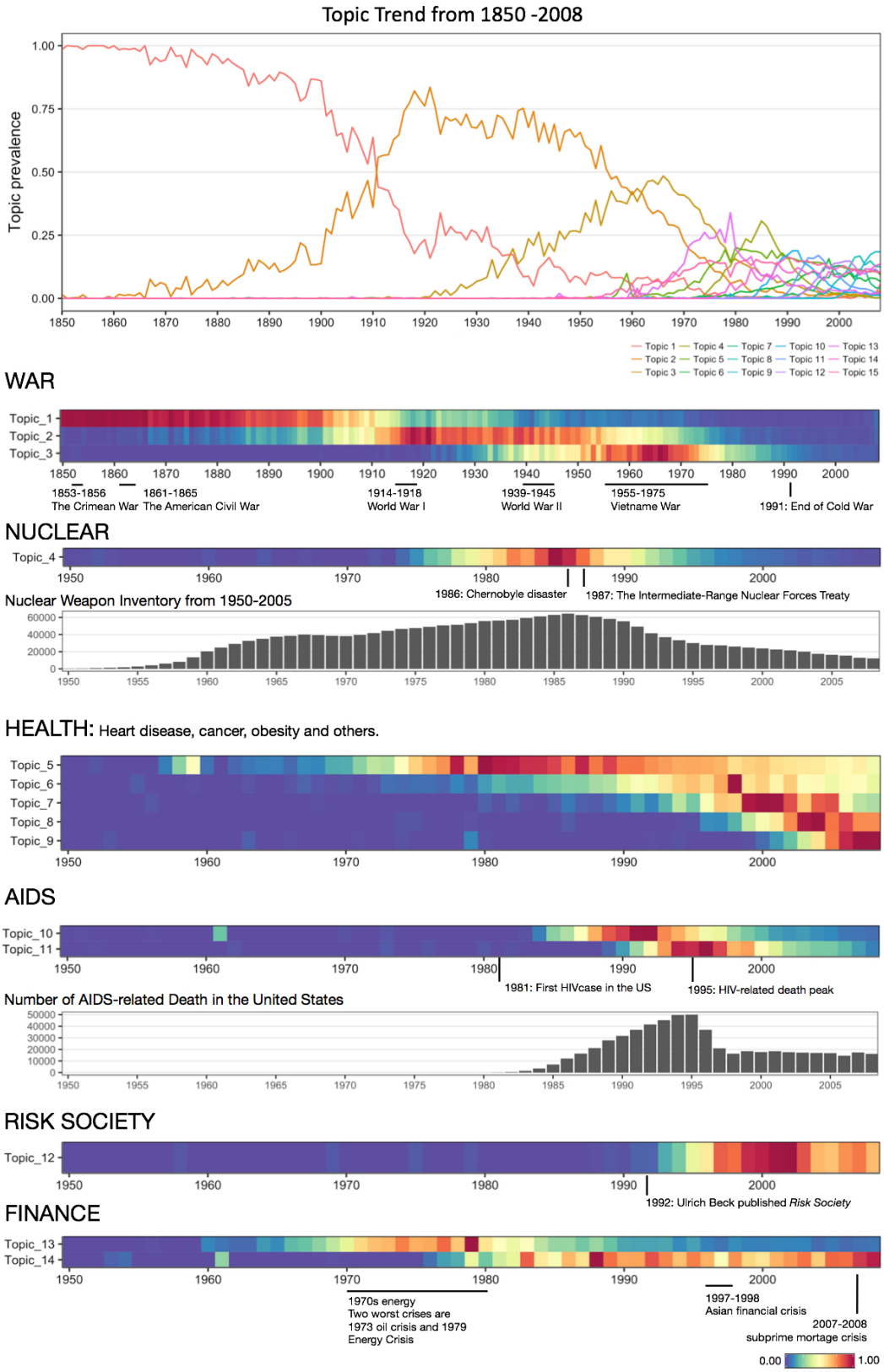


Figure 8.4. Trend analysis on risk topics derived from the Google Books Ngram Corpus. Topics are grouped into six categories: war, nuclear, health, HIV/AIDS, risk society, and economy. Relevant historical events are labeled to indicate how changes in the meanings of risk were associated with historical events and developments. Top panel: historical trends of 15 risk topics (computed using Equation 3). Bottom panel: normalized topic trend for each individual topic. Topic 15 is not included as it does not refer to a specific risk topic.

4.5 Discussion

Risk is a complex multivariate construct. It takes a variety of forms in public discourse and has, accordingly, been investigated in various ways. Each approach focuses on some dimensions of the discourse at the expense of others. One common approach has been to analyze media coverage of risk as a leading source of information for the general public and experts alike (see, e.g., Coombs and Slovic 1979, and various references in Young et al. 2008). Our approach consisted in a large-scale analysis of historical text corpora. Such corpora are attractive because they collate a vast array of perspectives on an extensive historical time window: in the case of the Google Book Ngrams Corpus, over 8 million books and 150 years. What did we learn about the risk-related discourse in English-speaking countries?

First, we found—consistent with Beck's (1992) diagnosis of post-industrialist Western societies as risk societies facing a wide variety of unique and human-made risks and with Gidden's (1990) idea that society is increasingly preoccupied with the future and its safety—that the word *risk* has become much more prevalent (Figure 1A), finding evidence of an approximately fourfold increase in its usage since the 1950s. Beck also stressed that risks in the post-modern world are increasing unknowable and unpredictable due to scientific and technological innovations having unanticipated consequences. It is possible that this process has contributed to our second major observation, namely, that the sentiments associated with risk have become much more negative, starting around 1900 and confirming Pinker's (2011) observation that humans have become increasingly preoccupied with the negative aspects of risk. Interestingly, the same does not apply to its close semantic relatives (Figure 1C). What is puzzling is that this change in sentiments is happening at a time when the semantics of risk have become increasingly associated with notions of quantification, reduction, and prevention—findings that also challenge the idea that the increase in negative sentiments has been caused by the unknowability of risks. In addition, we found that the risk categories to some extent reflect real-world changes in the prevalence and magnitude of the respective risks (see Figure 4 and our analyses of nuclear proliferation and AIDS-related deaths). Finally, we also found a shift from macro-risks, such as war and battle, to more individual-specific, chronic risks such as disease (Holzmann and Jørgenson 2000) as well as shift toward more variability in risk topics. The strong focus on modern diseases challenges the view that people fear the wrong things (e.g., Renn 2014, Schröder 2018).

Many of these patterns observed are remarkable in part because they are monotonic: the notable increase in the frequency and negativity of the risk construct, and the increase in number of topics it encompasses. These changes are perhaps related to one another. One

potential underlying mechanism is the social amplification of risk (Kasperson et al. 1988, Moussaid et al. 2015, Jagiello and Hills 2018): the observation that, as information is transferred from one individual to another, people tend to share the more negative aspects of a risk at the expense of potential gains. In Jagiello and Hills (2018), an individual exposed to a balanced argument on nuclear power shared that information with another individual. As information was communicated from one individual to the next, the focus shifted increasingly to the downsides of nuclear power and away from its benefits. This pattern is consistent with the substantial evidence that negative information has more influence on decision making than positive information (Ito et al. 1998, Baumeister et al. 2001, Rozin and Royzman 2001). A second, related factor is that this effect may be further amplified by increasing communication over the period of our analysis. As Herbert Simon (1971) noted, "a wealth of information creates a poverty of attention" (pp. 40–41). With the unprecedented amounts of information now available, all other things being equal, the absolute amount of negative information has increased. In this environment, information that is better at being received, remembered, and reproduced has a selective advantage (Hills 2019). Because negative information clearly has an advantage in the marketplace of information, one may indeed expect a rise in the negativity of the risk construct.

What is the state of the public discourse on risk? Our analysis can offer only a glimpse of this complex and multi-dimensional construct. We found results that were both disconcerting and reassuring. Primarily, the increasing prevalence of the word *risk* is an indicator of its growing significance, which is in itself a double-edged sword. Classifying something as a potential risk is likely to burden it with negative sentiments. Yet, branding something a risk also appears to imply the chance of a positive change in our fortunes. Importantly, the text corpus analyses suggest that risk categories track real threats over the 20th and 21st century, shifting from violent death to chronic disease. In this sense, the risk discourse reflects humans' changing relationship with threats and the plausibility of preventing them.

Chapter 5 : Words to Linguistic History

The Macroscope: A Tool to Examining the Historical Structure of Language

The recent rise in digitized historical text has made it possible to quantitatively study our psychological past. This involves understanding changes in what words meant, how words were used, and how these may have responded to changes in the environment such as healthcare, wealth disparity, and war. Here we make available a tool, the Macroscope, for studying historical changes in language over the last two centuries. The Macroscope uses over 155 billion words of historical text, which is growing as we include new historical corpora, and derives word properties from frequency of usage and co-occurrence patterns over time. Using co-occurrence patterns, the Macroscope can track changes in semantics, allowing researchers to identify semantically stable and unstable words in historical text, provide quantitative information about changes in a word's valence, arousal, and concreteness, as well as information about new properties such as semantic drift. The Macroscope provides information about both local and global properties of words, as well as information about how these properties change over time, allowing researchers to visualize and download data to make inferences about historical psychology. Although quantitative historical psychology represents a largely new field of study, we see this work as complementing a wealth of other historical investigations, offering new insights and new approaches to understanding existing theory. The Macroscope is available online at: http://www.macroscope.tech.

5. 1 Introduction

Hartley (1953) once wrote that "The past is a foreign country: They do things differently there" . Understanding why they did those things and what they were thinking when they did them is partly about history, but it is also falls under the umbrella of historical psychology. A number of recent accounts have documented apparent historical changes in the way people thought in the past. These accounts follow in the footsteps of well-documented historical changes that have taken place even in the last several centuries, for example, in the diffusion of print materials and the industrial revolution's disarming of the Malthusian trap, releasing large parts of the world's population from hand-to-mouth economies (Clark, 2008; Eisenstein, 1980). These changes have led to numerous claims explaining the rising spectre of

risk in society (Beck, 1992), the whittling away of violent behavior by the civilizing process (Pinker, 2011), urbanization's empowering of individuality and materialism (Greenfield, 2013), and the evolution of American English in response to information crowding (Hills & Adelman, 2015). The growing consensus appears to be that historical data represents a fertile ground for rolling our contemporary understanding of psychology back into the past.

The most common approach to studying historical beliefs and attitudes is what historians and literary critiques call *close reading*. A close read involves a human reader, who reads over original texts, attending to individual words and sentences. Scaling this approach to the volume of historical text currently available to make broad quantitative generalizations at the scale of hundreds of years is effectively impossible. A person reading 50,000 words a day would require 22,000 years to close read the text currently available in Google Ngrams book corpus. Over the past several decades, however, cognitive and language scientists have developed computational tools for *distant reading*, where researchers use algorithms to extract meaning from billions of words of text. These have been used to study properties of word recognition (Jones & Mewhort, 2007), the structure of memory (Hills, Jones, & Todd, 2012), the relationship between natural language production and individual differences (Pennebaker & Stone, 2003), changing frequencies of word usage across individual lifespans (Le, Lancashire, Hirst, & Jokel, 2011), and changes in word use over hundreds of years (Michel et al., 2011). In doing so, this progression has moved language analysis from synchronic investigation of single words to diachronic investigations of texts across cultural time, all of which can take place in the lifetime of a single researcher (or even in an afternoon).

The goal of the present work is to introduce a tool that adds an additional layer of structural depth to quantitative historical analysis, allowing researchers to zoom in and out on words--specifically, their semantics, and the associations they maintained in historical language. We call this tool the Macroscope, after the device in Piers Anthony's (1974) book by the same name which could zoom in and out on the cultural history of other alien civilizations. The key conceptual assumption upon which the Macroscope stands is that words provide information about the past and we can infer the meanings of those words through the relations they keep with other words (e.g., Firth, 1957). Thus, meaning is derived through historical context, providing a new way of looking at semantic history. In what follows we describe the underlying computational machinery of the Macroscope and provide several case studies that demonstrate the Macroscope's utility for understanding historical language.

5.2 Method

The Macroscope is a user interface consisting of a client-server interaction. The server, built in Node.js, handles user queries and analyses them in real time using Python. The data is then visualised on the client's website. It can be found at: http://www.macroscope.tech.

The Macroscope takes as input specific words of interest from the user, examines these in relation to a language corpus provided by the Macroscope, and outputs a range historical indicators about changing semantics over time. Here we take semantics in the broadest possible sense (see below). Data for each historical indicator can be downloaded in .csv format to the user's computer. A representation of the online interface for the Macroscope is shown in Figure 5.1. The details of the language corpora and computational algorithms are provided below.

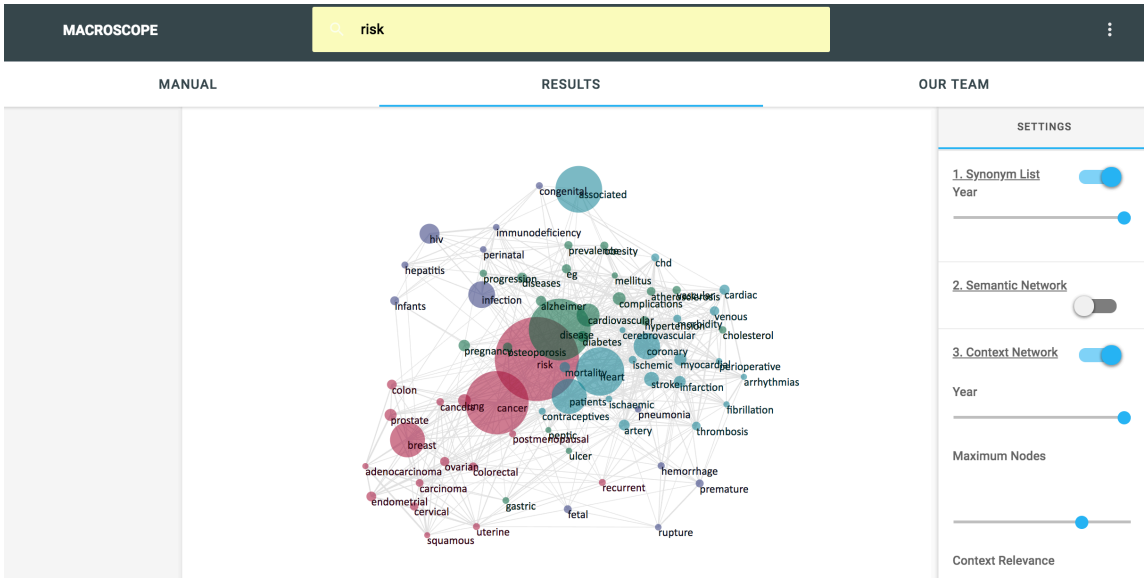


Figure 5.1. Screenshot of the Macroscope website. The search bar is on the top where users can input word of interest (state in the figure). The control panel on the right allows selecting specific analysis and manipulating parameters.

The language corpora. The first iteration of the Macroscope uses text from the English Google Ngram Book corpus (5-grams) (Michel et al., 2013). This will be supplemented with additional corpora (such as the Financial Times corpus and the Corpus of Historical American English) in forthcoming iterations, allowing users to compare data across multiple corpora. The Google Ngram Book corpus represents ~4% of all books published over the last several hundred years (Michel et al., 2013). Because the data representation is fairly sparse prior to 1800, we present data from 1800 to 2009 which contains approximately 155 billion words.

Frequency. Usage frequency is computed by dividing the number of instances of a word in a given year by the total number of words in the corpus in that year. For instance, in 1861, the word *slavery* appeared in the corpus 21,460 times, in 11,687 pages of 1,208 books. The corpus contains 386,434,758 words from 1861; thus the usage frequency of *slavery* in

1861 is 5.5x10-5. Users can input a search term into the search field and adjust various settings to capture and visualize the data of interest.

Co-occurrence matrix. To compute word properties from the words that a given word co-occurs with, the Macroscope relies on co-occurrence. The Google Ngram data consists of a matrix using 5-gram data. The matrix records the number of times any two words co-occurred within a 5-gram over 209 years from 1800 to 2009. We include the top 50,000 most frequently used words across the 209 years, resulting in a 50,000 x 50,000 x 209 matrix. Each word in the co-occurrence matrix is represented as a vector of dimension 50,000 that stores its contextual information.

Sentiment and concreteness. Using the co-occurrence matrix, the Macroscope computes contextual sentiment (valence), arousal, and concreteness by taking the mean of the relevant ratings of all the words that co-occurred with a given word in a given year. We used the Warriner, Kuperman, and Brysbaert's (2013) norms to retrieve contemporary valence and arousal ratings for each word, and the Brysbaert, Warriner and Kuperman's (2014) norm to retrieve contemporary concreteness ratings for each word.

Diachronic word embeddings. To find out which words are most semantically similar to each other and quantify their degree of similarity, we used distributional semantics, in which words are embedded in vector space according to their co-occurrence relationships (Bullinaria & Levy, 2007; Turney & Pantel, 2010). We constructed diachronic word embeddings for each year to allow comparisons across different years. This approach has been effectively demonstrated in a number of studies (Sagi et al., 2011; Xu & Kemp, 2015; Hamilton et al., 2016)**.** In our study, we constructed word embeddings as follows. First, vectors containing the number of times a given word co-occurred with all other words were directly obtained from the co-occurrence matrix described above. Second, we computed Positive Pointwise Mutual Information (PPMI) for each pair of words and constructed a PPMI matrix with entries given by

$$\mathrm{PPMI}(v_i, v_j) = \max(0, \log(\frac{P(v_i, v_j)}{P(v_i) \times P(v_j)}))$$

where v_i, v_j represents a pair of words from the corpus. *p(v)* corresponds to the empirical probabilities of word co-occurrences within a sliding window size of 5 over original text. Comparing to co-occurrence count, PPMI penalises importance of high-frequency words (i.e., *of, the, and*) that were used in the same context with a wide range of words, and favours words that frequently appeared together but not with others (i.e., *hong* and *kong*). Forcing PPMI values to be above zero ensures that they remain finite and this has been shown to improve

results (Bullinaria & Levy, 2007; Levy, Goldberg & Dagan, 2015). Lastly, we reduced the dimension of word embeddings to 300 using Singular Value Decomposition (SVD). The dimensionality reduction acts as a form of regularization and allows us to compare word similarities by computing cosine similarity of word embeddings.

To validate that the word embeddings we trained on the Google Ngram corpus accurately capture semantic relationships among the words, we tested these embeddings on 200 multiple-choice synonym questions collected by Levy, Bullinaria, and McCormick (2017). Each question corresponds to a set of five words: the test word, followed by the correct synonym, followed by three incorrect choices. Because some of the low-frequency words (such as consommé and treacle) were not included in our analysis, we tested 183 synonym questions using word embeddings trained on aggregated data from 2000 to 2008. Our performance (89.5% correct) was comparable to that of word embeddings trained using five different algorithms by Levy and his colleagues (accuracy rates ranging from 86.5% to 92.0%).

5.3 Results

Quantifying Semantic and Contextual Change. The Macroscope provides researchers with the ability to examine two distinct but related aspects of linguistic change in individual words over historical time as shown in Figure 5.2 below. First, diachronic word embeddings computed from the co-occurrence matrix enable us to discover words that are semantically similar to a given word for a given year (i.e., the semantic or synonym structure surrounding a word). These semantically related words are referred to as *synonyms* for the remainder of this paper (top half of Fig.5.2). Second, the co-occurrence matrix provides information regarding the context of a given word at a given year. Words that co-occur with the target word are referred to as *context words* for the remainder of this paper (bottom half of Fig.5.2).

On top of being able to "focus" the Macroscope on the semantics and contextual structure of an individual word in a particular year, the true power of the Macroscope is harnessed when the researcher "zooms" out to obtain a bird's eye view of changes in the semantic and contextual structure of words over historical time. Below we describe how the Macroscope can be used to examine the semantic (synonym) and contextual (co-occurrence) structure of individual words for a specific year (i.e., zooming in) and over historical time (i.e., zooming out). In the analyses described below, techniques from network analysis are employed to help with the interpretation and visualization of the synonym and co-occurrence structure of

words. All analyses can be easily replicated using the Macroscope and the user can download the network graphs along with the data used to construct the graphs.

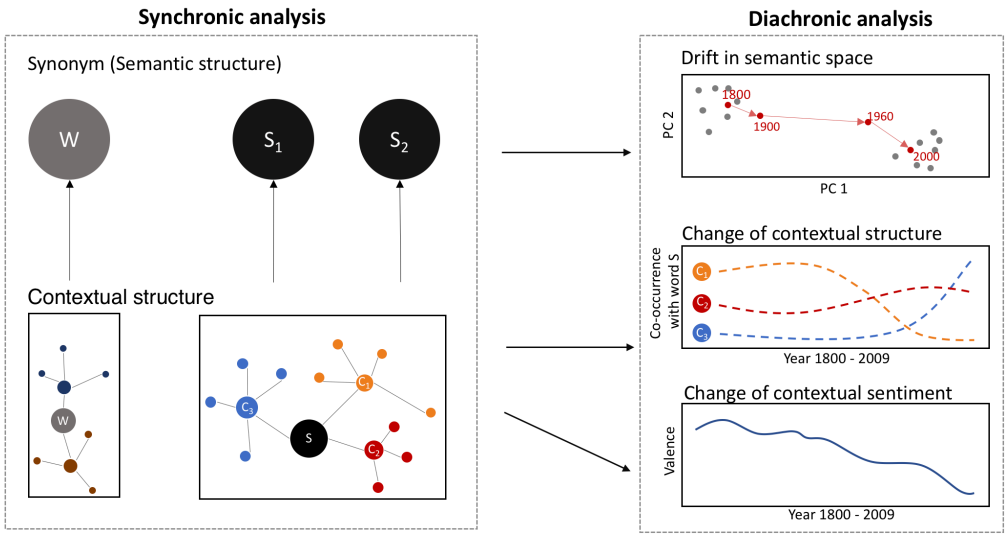


Figure 5.2. Conceptual framework summarizing the key features of the Macroscope. The Macroscope permits synchronic (left side) and diachronic (right side) analysis of the semantic/synonym (top) and contextual/co-occurrence (bottom) structure of words.

Synchronic semantic structure of words: Historical synonyms. How do we know what a word meant in the past? Using diachronic word embeddings, the Macroscope can quantify semantic similarity by computing the cosine distance of word embeddings for any pair of words. Therefore, a word's historical meaning can be inferred by finding its most semantically similar words in a given time period (i.e., *synonyms*).

Anxiety and depression are conceptualized as two distinct emotions by psychologists, yet often they are experienced by the general population as the same feeling (Barrett, 2017). To examine how these concepts are represented in the written language and produced and read by people who do not necessarily have a psychology background, we used the Macroscope to identify the synonyms of *anxiety, depression*, and *fear* using co-occurrence data from the year 2000 (see Table 5.1). *Anxiety* and *depression* share many synonyms that are associated with mental disorders. In contrast, *fear*, another commonly experienced negative emotion, appears to have different synonyms from *anxiety* and *depression*.

Table 5.1 Top five closest synonyms of depression, anxiety, fear, disgust, and anger from the year 2000, provided by the Macroscope.

Depression	Anxiety, Psychosis, Depressive, Hyperactivity, Disorder
Anxiety	Depression, Mood, Paranoia, Panic, Ideation
Fear	Dread, Shame, Anger, Remorse, Despair
Disgust	Loathing, Dismay, Disappointment, Revulsion, Sadness
Anger	Resentment, Bitterness, Jealousy, Rage, Indignation

To better capture how these three emotion concepts are related to each other, the Macroscope provides a network graph representing the semantic similarity structure of their synonyms. The nodes shown in the network represent the top five synonyms for *fear, depression, and anxiety* as identified above, as well as the words *fear, depression*, and *anxiety* themselves. The edges between nodes are weighted by the strength of semantic similarity between word pairs (i.e., the cosine similarity between word embeddings). Edges that are greater than a threshold of .8 are shown in the network (this value can be set by the user). If the synonyms of two words share a high degree of semantic similarity (i.e., if they are connected to each other in the semantic network), this indicates that the two words are likely to be used in similar contexts and are semantically "close" to each other. Higher semantic similarity among the synonyms of two words offers an additional layer of depth to investigate how similar are the meanings of the two words, even if the synonyms of the two words were not necessarily the same. Though previous tools have provided quantitative information about word similarity (e.g., BEAGLE from Jones & Mewhort, 2007; LSA from Landauer, Foltz, & Laham, 1998), the present example demonstrates how the Macroscope provides and visualizes additional information about the broader semantic similarity structure of words via their synonyms. Figure 5.3 (left panel) shows that the synonyms of anxiety and depression are synonyms of each other but are distinct from those of fear. Although psychologists treat anxiety and depression as two separate constructs, they appear to be used in semantically similar contexts in written language.

The same network approach used to represent concepts and their synonyms can also provide insights into the overlapping and distinctive components of two concepts. A similar analysis was conducted for the emotion words *fear, disgust*, and *anger*, three of the six basic emotions that are proposed to exist universally across cultures (Ekman, 1992). The results indicated that all three negative emotions intersect with some of each other's synonyms (see Table 5.1). Figure 5.3 (right panel) shows that the concepts of anger, fear, and disgust share

similar connections to such words as *disappointment, bitterness*, and *loathing*. However, each of these emotion concepts is also marked by its own unique components, which make the concepts distinct from each other: *disgust* is linked with *dismay, anger* with *rage* and *resentment,* and *fear* with *dread* and *shame.*

Figure 5.3 (a) Left: Synonym structure of anxiety, depression, and fear. (b) Right: Synonym structure of disgust, fear, and anger. The size of nodes is proportional to their usage frequency in the year 2000. The nodes represent the emotion concepts of interest and the top 5 most similar synonyms for each of the emotion concepts. The colors represent the community structure of nodes in the network and each community is represented with a different color. Community structure was detected by algorithm proposed by Blondel, Guillaume, Guillaume and Lefebvre (2008).

Diachronic semantic structure of words: Semantic drift analysis. With large diachronic language data, the Macroscope is able to track how the semantics of individual words change over time. In the following examples we show how several words "move" along a path in a semantic space defined by their historical synonyms. A longer path moving from one point in the semantic space to another indicates significant changes in a word's semantic meaning over time. In contrast, a path that stays within a confined semantic space suggests that the word has retained its meaning over the time window examined.

Using the Macroscope the user can conduct a semantic drift analysis by inputting the word of interest, beginning and end time points (e.g., year 1850 and 2000), and intervening intervals (e.g., spaced by every 50 years). A semantic space was constructed for a target word by searching for its historical synonyms at the beginning time point (1850) and its modern synonyms at the end time point (2000). All synonyms' word embeddings are taken in their modern sense (2000). We also retrieved historical word embeddings of the target word for each

time point of interest (i.e., 1900, 1950) and align their historical embeddings to its modern embedding using orthogonal procrustes (Schönemann, 1966), an algorithm to map one matrix to another of same shape. Lastly, these word embeddings were visualized on a two-dimensional space using principal component analysis (PCA). All synonyms in this two-dimensional space are represented in their modern sense. Although in reality all word meanings fluctuate over time, we elected to adopt this approach in order to provide a clearer understanding of how changes in a word's historical meaning occur over time as benchmarked against its modern sense.

We used the Macroscope to examine the semantic change of three words that have been previously documented in historical linguistics (Jeffers & Lehiste, 1979). Figures 5.4a to 4c shows the semantic drift analysis of *broadcast, cell,* and *car* from the year 1850 to 2000 (with 50-year intervals). In 1850, the word *broadcast* referred to 'disperse upon ground by hand' and was closely associated with agricultural activity. In 2000, the word *broadcast* referred to radio and other media-related concepts. Our analysis shows that this change primarily took place between 1900 and 1950, a time period during which radio and television were invented (Fig 5.4a). *Cell* changed its dominant meaning from "a chamber in a prison" to a biological term and this change predominantly took place between 1850 and 1900 (Fig 5.4b). In 1850 the word *car* referred to a horse-driven wagon, but after the automobile was invented in 1885, it quickly acquired its modern sense. The semantic drift analysis shows that by 1900, *car* was no longer associated with a wagon (Fig 5.4c), but with modern transportation vehicles like *bus* and *truck*. In addition, we conducted a similar analysis for a word that was likely to be semantically stable over time: *happy*. The semantic drift analysis confirmed our intuitions: The word *happy* remained within the same semantic space over the past 150 years.

Semantic drift analysis shown in Figure 5.4 offers a qualitative visualization on how word meanings changed over history, but it is not easy to quantitatively compare semantic stability between words (i.e. the semantic path travelled by *happy* relative to path travelled by *broadcast* from 1850 to 2000). Previous work has examined the properties of words that appear to show the highest degree of stability over historical time (e.g., Pagel, Atkinson, & Meade, 2007; Monaghan, 2014, Hamilton et al, 2016). Since the Macroscope provides information on diachonic changes in semantics, it can be used to quantify semantic stability of words as shown above in Figure 5.4,

$$Stability^{t}(w_i) = \cos\_dist(w_i^{(T)}, w_i^{(T+t)})$$

where $w_i^{(t)}$ refers to the word embedding of word w_t in year t. Semantic similarity ranges from 0 to 1. For example, similarity of *happy* between year 1850 and 2000 is 0.74, much higher comparing to words underwent greater semantic change, such as *broadcast* (0.08), *cell* (0.17), and *car* (0.47). This allows researchers to examine potential forces that influenced semantic change. As a baseline for further examination, the Macroscope provides semantic stability of a word in relation to its modern and historical word embeddings. Use this method, we retrieved the 10 most stable words from 1800 to 2000. They are: *and, the, when, his, he, they, him, in, them, a*. A complete list of word stability between two time points can be downloaded from the Macroscope.

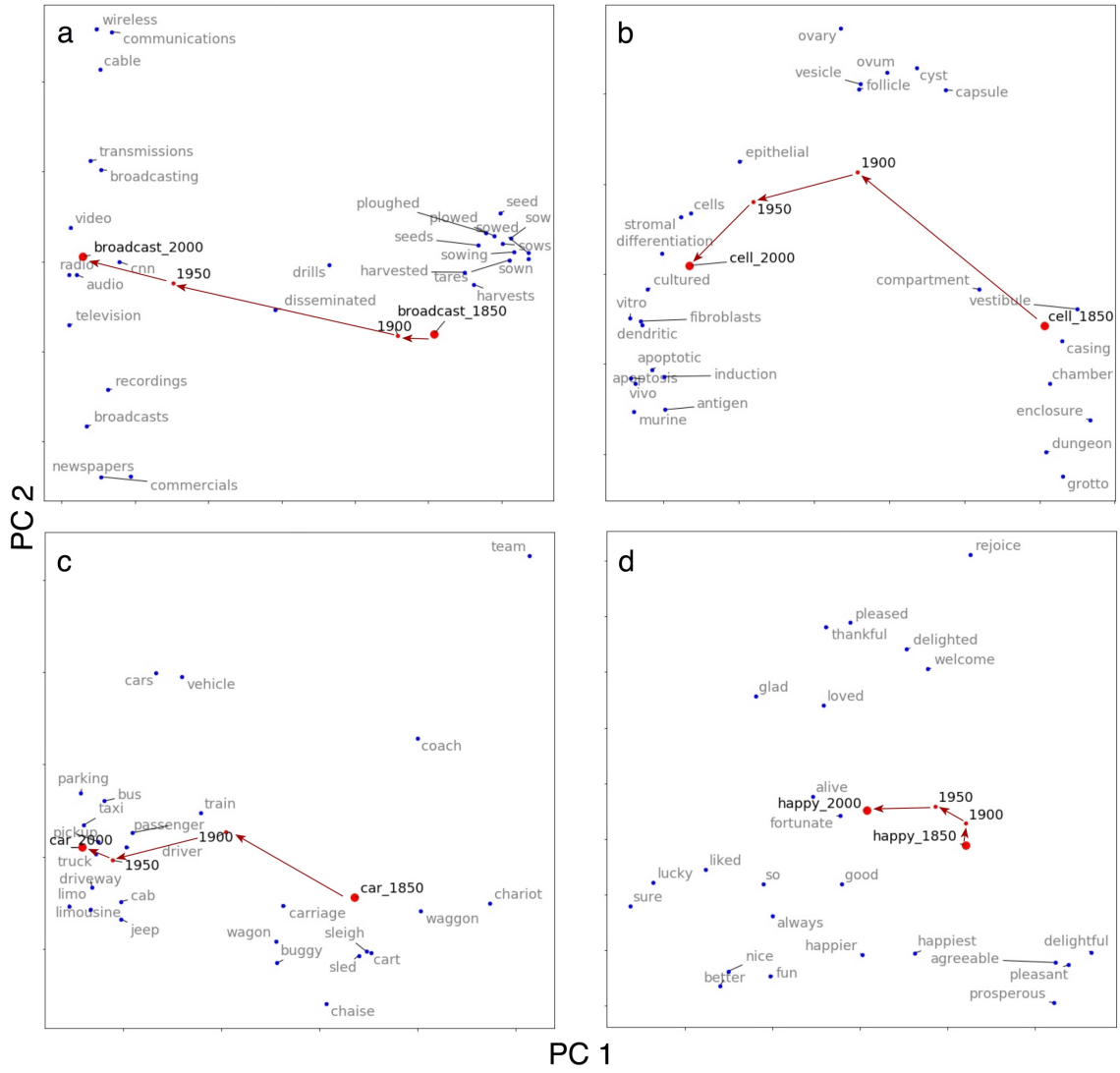


Figure 5.4. Semantic drift analysis for a) *broadcast*, b) *cell*, c) *car*, and d) *happy* from 1850 to 2000 with 50 year intervals. The blue dots indicate words that are semantically related to the target word of interest (i.e., its synonyms at the first and last time points). The path taken by the red dots indicate the "drift" in semantics of the target word from 1850 to 1900, from 1900 to 1950, and from 1950 to 2000.

Synchronic contextual structure of words. Synonym analysis provides an accessible way to examine the semantic structure of words based on the conceptual assumption that words that are used in similar contexts are also semantically related to each other (e.g., Jones & Mewhort, 2007). On the other hand, identifying the particular context(s) in which a word was used can help us understand how polysemous words are used in their different senses across varying contexts, furthering our understanding of the relationship between the semantic and co-occurrence structure of words. For instance, it is possible for words to have a stable semantic/synonym structure but a varying co-occurrence structure over time. A concrete example can be seen in the word *woman*. Although the semantic meaning of the word *woman* has not changed much over past 200 years, in recent decades the word *woman* has been increasingly used it in the context of social issues surrounding feminism, gender discrimination, and abortion--contexts that were not commonly discussed during the 1800s.

The following co-occurrence networks of the words *monitor*, *option* and *gay* shows how the Macroscope can be used to understand the contextual structure of words. All networks were centred at the target word of interest. The context words, represented as nodes in the network, were selected based on their Positive Pointwise Mutual Information (PPMI) value with the target word. The edges were weighted by the PPMI values between each word pair. Next, nodes with low co-occurrence frequency with the target word and edges signalling low PPMI values were removed. Lastly, nodes with no edges (i.e., isolates) are removed. During the procedure, arbitrary thresholds for parameters must be specified in order to produce meaningful network graphs. The networks presented below were constructed using a PMI threshold of 3, and a minimum co-occurrence frequency of 200 times per 10 billion words. Communities are sub-groupings of nodes that are more likely to be connected to each other than to other nodes within the network. Community structures of the network are detected using an algorithm introduced by Blondel et al (2008) based on modularity optimization that uses an iterative process which defines each node as a community at the first step and merges them until modularity (a measure of the strength of the communities) is optimized.

Figure 5.5a shows the contextual network structure of *monitor* in the year 2000. Community detection analysis of the contextual network showed approximately 3 distinct contexts in which the word was used: as a computer device, in healthcare related settings, and a group of verbs that it often accompanies. From the contextual network structure of *monitor*, one can infer that it was used as a noun or a verb. As a noun, *monitor* is often referred to as a computer device; as a verb, *monitor* is often used in medical settings.

Figure 5.5b shows the contextual network structure of *nuclear* in the year 2000, which shows that the word *nuclear* is used in a number of distinct contexts: It can refer to a power source, physics phenomena, a technology known as nuclear magnetic resonance (NMR), or a weapon associated with some countries (*Soviet, Cuba, Korea*) but not other nuclear-armed states.

Figure 5.5e is an example of what the contextual structure of a polysemous word such as *option* looks like. Other than the conventional meaning of choosing among various possibilities, *option* also refers to a financial instrument. As Figure 5.5e shows, its contextual structure in the year 2000 is divided into two components. One involves its traditional sense, which incorporates the use of the option button on a keyboard. The other component consists of finance-related terms. It is important to note that such information would not be available if one only analysed the synonyms of *option* in the year 2000 (which are *options, cancel, default, item,* and *choose*), further highlighting how an analysis of a word's contextual structure can complement the analysis of a word's semantic structure.

As mentioned earlier, understanding the contextual usage of a concept can be useful to infer changes in the sociocultural environment. Figure 5.5c shows the context in which the word *gay* was used in the year 2000. It was not only associated with homosexuality, but also with a political movement associated with issues that extended beyond gay rights, such as feminism and abortion. Sexually transmitted diseases such as HIV and AIDS also appeared in this context, reflecting a social awareness of the association between homosexuality and the way that these diseases were transmitted among communities of gay men during the AIDS epidemic in the 1980s and 1990s. In contrast, 150 years ago, not only did all these associations not exist, the word *gay* simply did not refer to homosexuality. The contextual structure analysis suggests the word *gay* in 1850 was used in contexts involving fashionable clothes, cheerful mood, and pleasant colours (Figure 5.5d).

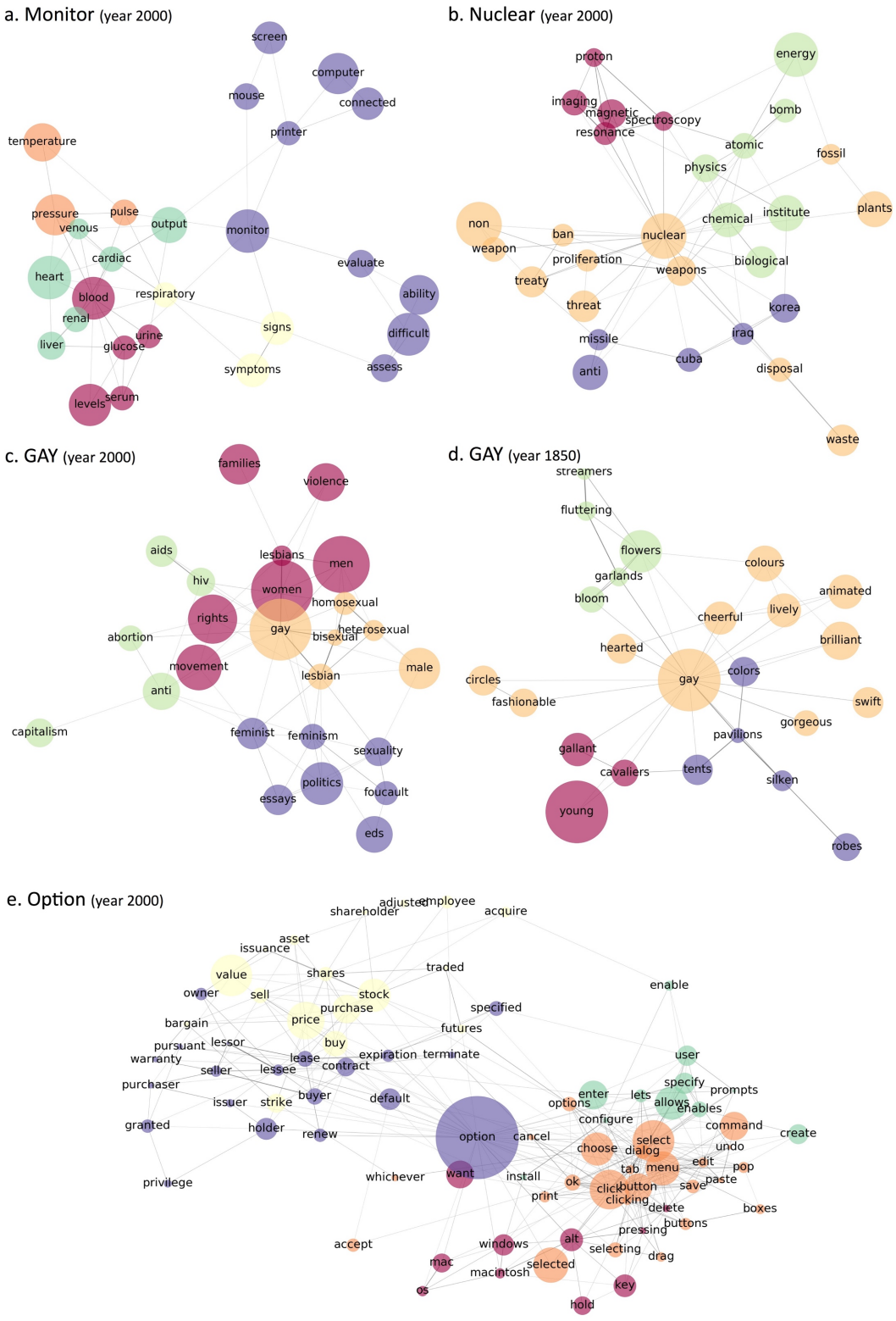


Figure 5.5. The contextual network structure of a) *monitor*, b) *nuclear*, c) *gay* in year 2000, d) *gay* in year 1850, and e) *option*. The nodes represent the context words that co-occurred with the target word in a given year. The size of nodes is proportional to their usage frequency in a given year. The nodes were included in the networks if they had a PMI threshold greater than 3 with other words, and a minimum co-occurrence frequency of 200 times out of 1 billion words with the target word. The colors represent the community structure of nodes in the network and each community is represented with a different color.

Diachronic contextual structure of words. In addition to quantifying the contextual structure of words at a static point in time, the Macroscope allows users to quantify changes in the contextual structure of words diachronically. Figure 5.6 below shows how the frequency of co-occurrence of words co-occurring with *gay* and *nuclear* have changed between the years 1950 and 2000. Words with larger blue bars to the right (top of the y-axis) are words whose frequency of co-occurrence with the given word has increased the most from 1950 to 2000, whereas words with larger red bars to the left (bottom of the y-axis) are words whose frequency of co-occurrence with the given word has declined the most from 1950 to 2000. For instance, for the word *gay*, *lesbian* and *bisexual* increased the most in their frequency of co-occurrence whereas *happy* and *hearted* decreased the most in their frequency of co-occurrence. For the word *nuclear*, *weapons* and *magnetic* increased the most in their frequency of co-occurrence whereas *molecule* and *spin* decreased the most in their frequency of co-occurrence, reflecting the increased usage of nuclear as a weapon of destruction in recent years as compared to its scientific sense in the 1950s.

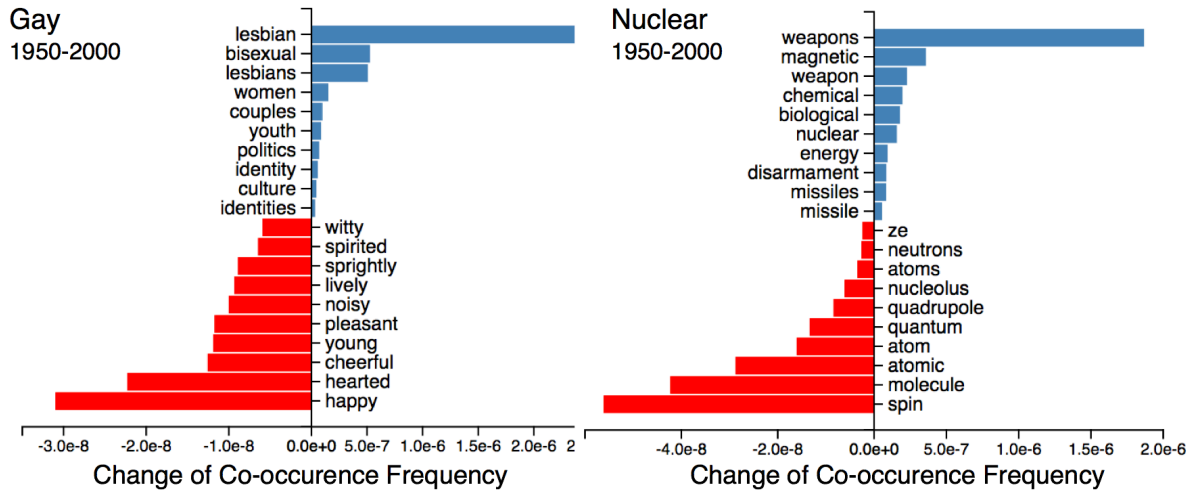


Figure 5.6. Words whose frequency of co-occurrence with *gay* and *nuclear* changed the most from 1950 to 2000. Words that increased the most in their frequency of co-occurrence with the target word from 1950 to 2000 are shown in blue near the top and words that decreased the most are shown in red near the bottom. The x-axes on the left and right side of the y-axis are scaled differently so that the y-axis is centered in the middle of the graph.

Although the previous analysis shows the largest changes in the frequency of co-occurring words between two time points, it is not completely clear to what extent a word has "lost" its old meaning. For instance it is possible for a word's old meaning to still be in use, albeit not as commonly used as before. In addition, the previous analysis does not contain information regarding fine-grained changes in the frequency of co-occurring words during the time period in between the two specified time points.

One way to address these questions is to examine the extent to which a given word co-occurred with words found in its historical context. These context words can be obtained from the synchronic contextual structure analysis described earlier (see Fig 5.5). Users of the Macroscope can also enter words of particular interest to their research. The co-occurrence value in Figure 5.7 below (on the y-axis) was computed by summing the number of times the target word co-occurred with each word of interest (in this case, from its historical context identified in the contextual structure analysis in Figure 5.5) in each consecutive year after the historical reference year.

For instance, *gay* in 1850 co-occurred with words associated with cheerfulness, bright colors, and fashion (Fig 5.5c) and in 2000 co-occurred with words associated with homosexuality and sexually transmitted diseases (Fig 5.5d). The Macroscope can take these two lists of context words and compute their respective co-occurrence frequencies with the target word *gay* to capture how frequently its meaning in 1850 and its meaning in 2000 were used over the entire corpus (i.e., from 1800 to 2009). Figure 5.7 (left side) shows that how overall usage frequency of *gay* can be largely decomposed into two trends, with each corresponding to a different sense of *gay*. The co-occurrence between *gay* and its context words in the year 1850 declined quickly after 1900, whereas the co-occurrence between *gay* and its context words in the year 2000 emerged in the mid-1960s and increased dramatically after the 1980s. The pattern suggests that the old meaning of *gay* has been largely overwritten by its new emerging meaning.

Another example is the word *option* (shown on the right side of Fig 5.7). When looking at the contemporary contextual structure of *option* (Fig 5.5e), one can easily see that the word *option* refers to economic instruments: A *stock option* refers to stock warranted from a company to their employees as part of a remuneration package and a *lease option* refers to a real estate contract that gives the lessor an option to buy the property. A visual inspection of Figures 5.7d and 5.7f shows that a lease option probably existed in some form before the 19th century whereas a stock option was first introduced in the 1920s and the usage of this sense continued to grow in the 1980s.

By combining the synchronic contextual structure analysis of words with a diachronic analysis of co-occurrence frequency of context words with the target word, the Macroscope provides an accessible quantitative approach to track the association strength between a word and its various contextual structures over history, which could be used to investigate the evolution of word meanings or cultural change over time.

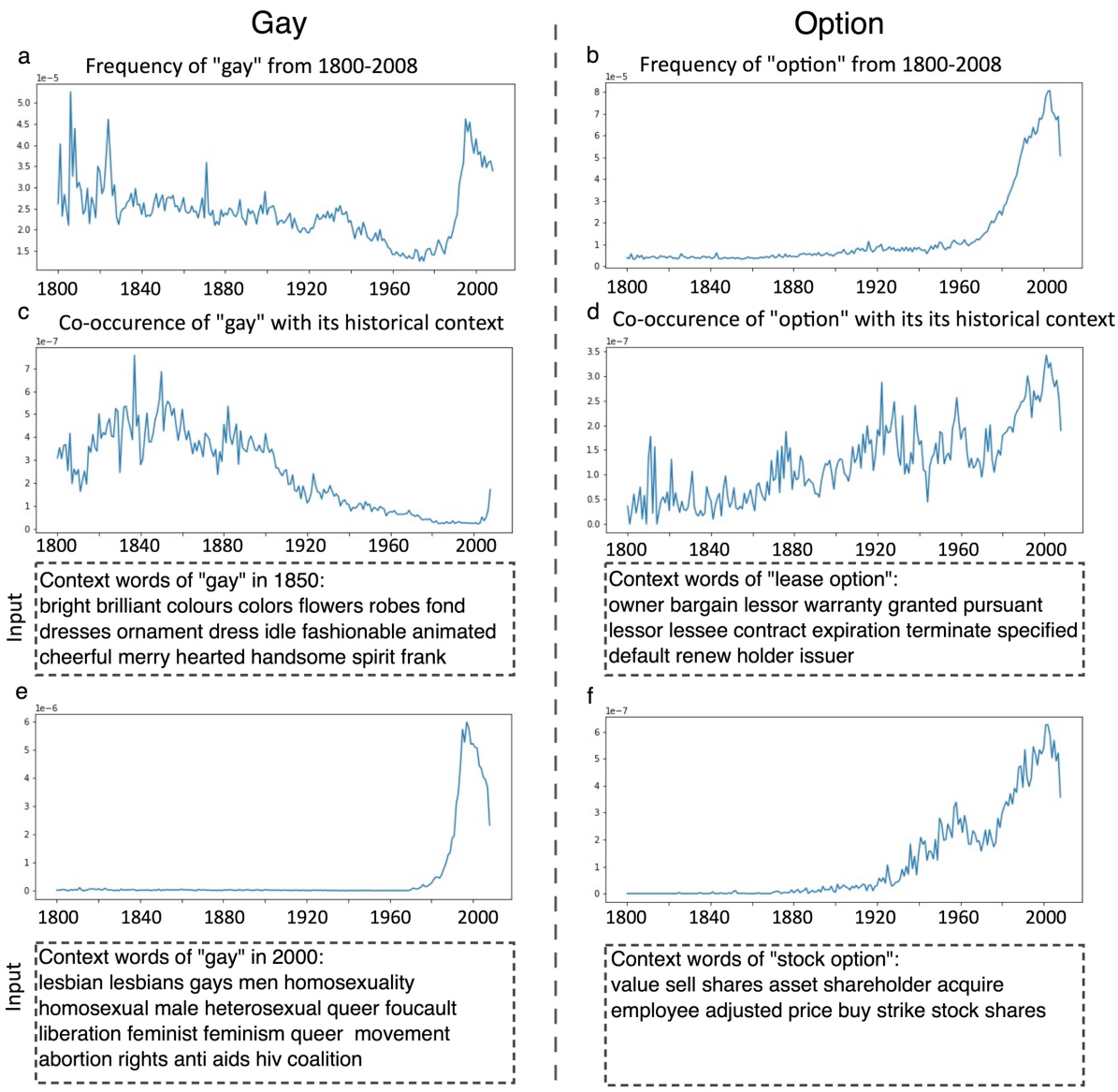


Figure 5.7. Co-occurrence frequency between the target word and its context words from 1850 and 2000. The context words were derived from the synchronic contextual structure analysis described earlier (see Figure 5.5 for examples). The co-occurrence frequency was computed by summing the number of times the target word co-occurred with each single word in the list of context words.

Diachronic changes in word sentiment. So far we have demonstrated how the Marcoscope can be used to investigate the semantic and contextual structure of words at a specific point of time and across historical time. Below we show how the Macroscope can also be used to examine diachronic changes in word *sentiment* and how that information can be used to infer cultural changes due to urbanization and understanding the changing social perceptions of risk.

Example 1: Cultural changes due to urbanization.

Greenfield (2013) analyzed the changing psychology of culture in the US as a consequence of urbanization by selecting two lists of words associated with urban and rural cultural values respectively and tracking their usage frequency over time. She found that words signaling urban values have proliferated in the US over the past century, along with a declining trend among words signaling rural values. The Macroscope can not only track the usage frequencies of these words over time, but also track the sentiment change of words over time. Here we use the Macroscope to extend Greenfield's results by analyzing the sentiment of words that co-occurred with words associated with urban and rural values over historical time.

The results reproduce Greenfield's analysis (see left side of Figure 5.8) showing that the frequency of *give* and *obliged* (rural values; in blue) decreased over time and the frequency of *get* and *choose* (urban values; in orange) increased over time. The Macroscope adds additional information by showing that the sentiment of *get* and *choose* increased at a faster rate as compared to the sentiment of *give* and *obliged* (see right side of Figure 5.8). The increasingly positive sentiment of urban value words compliments and extends Greenfield's argument because increasing usage of a word such as *get* and *choose* does not necessarily imply that urban values are viewed positively and are increasingly adopted by people. To provide a counterexample, if a word is used more frequently but has an increasingly negative sentiment (such as the word *gay* in the 1980s during the AIDS epidemic), this concept may instead be viewed as dangerous and unfavorable.

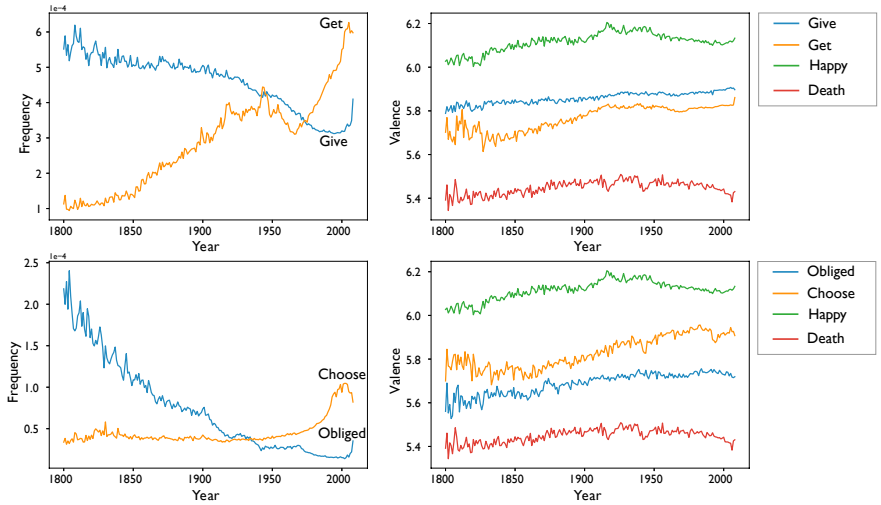


Figure 5.8. Frequency (left column) and valence (right column) from the Macroscope. The left side shows the usage frequencies for words associated with urban values (get and choose in orange) and words associated with rural values (give and obliged in blue) over historical time. The right graphs show the change in sentiment for the same words along with the change in sentiment for words such as happy and death respectively, a high- and a low- valenced word whose sentiment is stable over time.

Example 2: Changing social perceptions of risk.

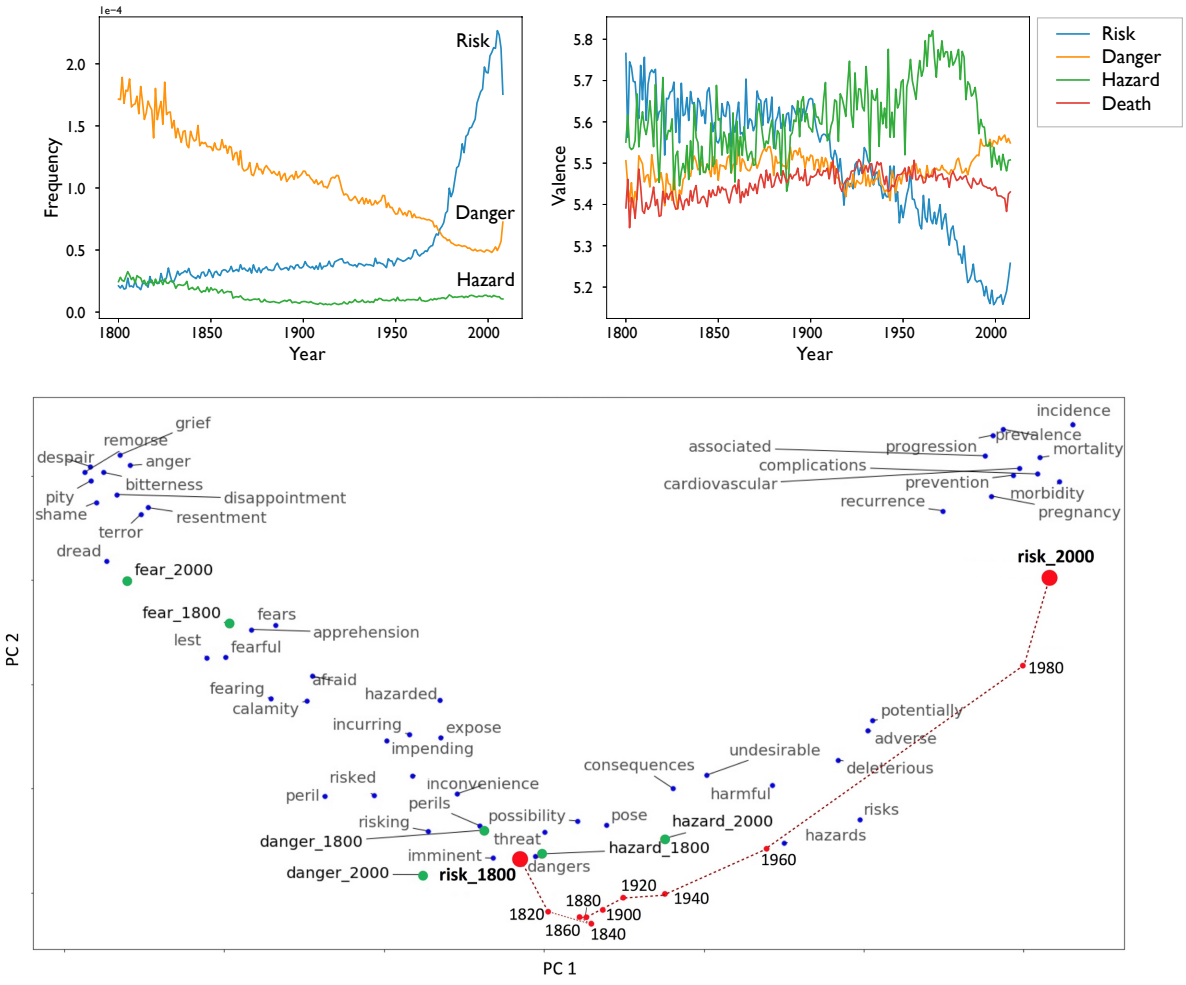


Figure 5.9 (a) Top left: Usage frequencies *of danger, hazard,* and *risk* over historical time. (b) Top right: Changes in the contextual sentiment of *risk, danger, hazard*, and *death* (*death* was selected as a benchmark) over historical time. (c) Bottom: Semantic drift of *danger, hazard*, and *risk* from 1800 to 2000. All figures were generated using the Macroscope.

Risk, as defined by the Oxford English Dictionary, is synonym for *danger*, *hazard* and *fear*. However, sociologists and anthropologists have argued that *risk* represents more than just objective dangers or hazards in the real world. Instead, the notion of *risk* has been used to motivate social regulation and control or acts as a surrogate for other ideological concerns (Berk, 1992). In this example, we used the Macroscope to examine the relationships between *risk* and its synonyms over the past 200 years. Our results show that *risk* usage has experienced a rapid proliferation after 1950s compared to a stable usage of *hazard* and a declining usage of *danger* (Fig 5.9a). Correspondingly, the contextual sentiment of *danger* and *hazard* remained stable over time whereas the sentiment of *risk* became increasingly negative (Fig 5.9b). Output from the Macroscope (Fig 5.9c) shows how *risk* and its synonyms (i.e., *danger* and *hazard*)

drift in semantic space between 1800 and 2000: *danger* and *hazard* have fairly limited semantic drift as compared to risk, which in the year 2000 was primarily associated with words related to medicine and health.

5.4 Discussion

Language has changed over historical time and that change is reflective of the kinds of things that people experienced and believed. The goal of the present paper is to introduce the features of the Macroscope, an online algorithmic tool for zooming in and out on the semantic and contextual structure of words across historical time. A key conceptual assumption that the Macroscope neatly capitalizes is that words provide information about the past and we can infer the meanings of those words through the relations they keep with other words. To summarize, the Macroscope can provide (i) synchronic and diachronic analysis of a word's semantic structure (based on word embeddings derived from the co-occurrence matrix), (ii) synchronic and diachronic analysis of a word's contextual structure (based on word co-occurrences), and (iii) diachronic analysis of a word's sentiment.

In the numerous examples presented above, we provide evidence that the meanings of words can be derived through its historical context in language, and this provides researchers with a new way of looking at semantic history through historical language. Importantly, these analyses can be easily conducted by anyone via the Macroscope, which can accessed online.

The Macroscope offers numerous inroads to investigating many contemporary problems in psychology and historical linguistics (e.g., Ladd, Roberts, & Dediu, 2015). For example, what properties of words influence semantic shift (e.g., Zalizniak, 2012). How do word senses changed over time in relation to other word properties such as frequency, concreteness, and age of acquisition (e.g., Ferrer-i-Cancho & Vitevitch, 2017; Monaghan, 2014; Zipf, 1946)? Can we use nowcasting methods to 'backcast', examining how word usage reflects the influence of historical events (Lampos & Cristianini, 2012; Hills, Proto, & Sgroi, 2015)? What are additional structural properties of language that are associated the birth and death processes of words (Pagel et al., 2007; Vejdemo & Hörberg, 2016)? To what extent have words used in studies of age-related cognitive decline changed during the lifetime of individuals under study, for example in studies of memory and association (Hills, Mata, Wilke, & Samanez-Larkin, 2013; Ramscar, Hendrix, Shaoul, Milin, & Baayen, 2014)? We feel this is the tip of a large iceberg of potential questions.

Historical studies of any kind are limited in their generality by the artifacts that survive, who originally produced them, and who they were produced for. Studies of historical language are no different (see Hills & Adelman, 2015). Thus, the Macroscope is naturally limited in what it can see. As far as we know, there are no spoken language corpora, which means that individuals who could not write will not be reflected (probably ever) in historical language analysis. Historical texts may have also focused on different topics over time and therefore may not offer usage patterns that reflect common topical environments. Better understanding these patterns and their consequences for language is part of the question we hope the Macroscope can answer. For example, Dubossarsky, De Deyne, and Hills (2017) showed that free association networks changed non-linearly across the lifespan, between the ages of 8 and 80. This is mostly likely due to both developmental changes associated with factors underlying human cognition and changes in the lexical environment since roughly the 1920s. What language corpora best reflects this changing population? It is difficult to say. But studies of historical language corpora nonetheless offer inroads into understanding what language structure can explain in the absence of additional assumptions. In forthcoming iterations of the Macroscope, additional corpora will be included to allow researchers to address specific question about generality.

To conclude, the language people use over historical time has been a primary source of understanding people's past beliefs and attitudes. The Macroscope brings quantitative approaches to a broader range of researchers interested in understanding historical psychology through the lens of language, enabling them to test and develop hypotheses about specific patterns of word usage and its semantics across history. In other words, the Macroscope is a passport to visit the foreign country of the past.

Chapter 6 Conclusions

Summary

This thesis set out to explore how words reveal psychology through three studies: development of a recall-based emotion scale (ERT), public attitudes towards immigrant groups, and a cultural history of risk. We demonstrated how quantitative text analysis can be applied to questions traditionally studied in either experimental paradigms or case studies. In this section, I will first summarise each study, then discuss their implications and limitations, and lastly delineate how the Macroscope may extend into future researches.

We first showed in chapter 2 that individuals' language production accurately reflects their emotional states. We proposed a new emotion scale named Emotion Recall Task (ERT) to measure emotion based on the 10 words participants produce to describe their feelings. This makes the ERT different from all existing affect scales that typically rely on recognition of emotion terms from a predetermined emotion checklist. Therefore, the ERT captures the breadth and specificity of emotions that are not available in other scales but that are nonetheless commonly reported as experienced emotions. We also showed ERT is reliable in a test-retest paradigm and strongly correlated with related constructs such as PANAS, various well-being scales, depression and anxiety.

Next in chapter 3, we scale up text analysis from individual level to group level in an analysis of social perceptions towards immigrant groups in the United States. We quantified historical change in language around 56 immigrant groups from a corpus that contains 20 years of news articles published on the New York Times. This is quantified in relation to sentiment, concreteness (a proxy for perceived social distance) and 15 immigrant-related topics. We found positive sentiment is strongly correlated with concrete descriptions, with concrete language predicting future positivity but not vice versa. Topic modelling reveals that public perception towards immigrants is complex and multi-faced. It identifies what topics drive the perceived positivity of each immigrant group. Together, it shows when large and representative sample of documents produced in a society is available, it can be used to infer social attitudes towards specific issues by investigating a relevant subset of the corpus.

In chapter 4, we extend our scope from a synchronic view on culture phenomena to diachronic analysis of culture change. We reconstructed what has been perceived as a risk over different historical periods. Using two large corpora, the Google Books Ngram Corpus and The New York Times Annotated Corpus (NYT Corpus), we studied the historical dynamics of the conceptualization of risk by analyzing language used to construct and single out risk. We found

that *risk*, unlike its synonyms *danger* and *hazard*, over the past two centuries has undergone tremendous semantic change, used increasingly frequently and appeared in more negative contexts. In terms of risk topics, public attention to risk has shifted from war to chronic health and threats to economy. These results may further inform future public discourse around risk.

In Chapter 5 we introduce the Macroscope, a linguistic tool that examines historical language structure. Using co-occurrence statistics derived from Google Ngram Corpus, the Macroscope offers 2 synchronic analysis – identifying synonyms and contextual structure of a given word – and 3 diachronic analysis – historical change of a given word in terms of its contextual sentiment, semantic drift, and co-occurrence frequency with other selected words. The Macroscope provides information about both local and global properties of words, allowing researchers to visualize data to make inferences about historical psychology.

Implications to emotion measurement

Since the 10 emotion items on the ERT scale are generated by participants, the ERT scale received by one participant is essentially different from others. Therefore, some may critique this leads to comparison between "apples and oranges". However, we must acknowledge that emotional experience is high-dimensional, complex, idiosyncratic and often not comparable between one and another. Moreover, even though a direct comparison of an apple and an orange is not possible, one can certainly compare their weight, colour, size, or other shared features. Similarly, it is not the holistic representation of emotion experience the ERT tries to capture and compare. Instead, the ERT extracts one dimension, the positive-negative affect, from recalled emotion experiences of different individuals, and compares them on this common dimension. Like a Swiss army knife that flexibly adapts to task at hand, the ERT, by inviting participants to construct the emotion scale, adapts itself to capture emotion space of different degree of complexity.

The ERT offers more information than averaged valence of the 10 words produced in the task. The semantic meaning of the recalled words, recall sequence, and reaction time may inform other emotion-related phenomena, such as emotion intelligence (Salovey & Mayer, 1990), emotion granularity (Tugade, Fredrickson, Barrett, 2004), and pattern of memory search in the emotion space. For example, some of our preliminary analysis suggests that participants who can better elaborate their emotions (produce less gap between recalled emotions and recognised emotions) score higher on emotion intelligence. The words produced in the ERT may also be used to infer emotions that are not explicitly reported: instead of directly asking a

person how anxious he is, we can infer his anxiety through semantic similarity between the 10 ERT words and *anxiety*.

The ERT may also be used as a potential paradigm to study the representation of emotion memory. Beside from what has been reported in Chapter 2, I conducted interviews on 10 participants asking them "what comes up to your mind before you generated each of the 10 words that you have used to describe your emotion". The answers reveal that words produced in the ERT could be used as a surrogate for highly diverse emotional-laden ideas. The ideas preceding the production of words could be very detailed such as *family party with parents, boyfriend, and cousins last Sunday at home* (word produced is *happy*), or highly symbolic like *lying in bed alone in a dark bedroom for hours* (word produced is *wasted*), or abstract images such as *bright blue dotes* (word produced is *holiday*). Future research can investigate the diversity of emotion representations in mind.

Implications to culture studies

Common critiques to quantitative analysis of corpus often centre at the representativeness of the corpus: exactly whose voice has been included? Pechenick and his colleges have called for caution in using Google Ngram Corpus in cultural studies because it is overly represented by academic articles (Pechenick, Danforth & Dodds, 2015). Moreover, concerns have been raised on whether Google Ngram Corpus only reflects the culture among a small proportion of the population who has access to publication. Moreover, instead of reflecting cultural dynamics, language change over history may be the result of the spread of literacy and publication privilege to wider range of social groups.

I agree that these limitations must be analysed before making any conclusions. However, I disagree that a corpus that representatively include language produced by all members in a society is the best one to infer culture. After all, culture is not a democratic summation of minds across every social member. Instead, cultures, especially national cultures, resonate with the voices of the powerful, and are filled with the silence of the powerless majority (Kramsche, 2009). Influential people are more likely to publish their ideas, and through the diffusion of their published work their ideas become even more influential. The Google Ngram Book's ignoring of the voice from the silent majority (who rarely turn their ideas into published works) may sometimes be used as an asset instead of a liability.

Another potential issue is that since results from corpus analysis often concern historical cultural change, experimental or empirical validation is not always possible. It must be kept in mind that a quantitative approach might lead to omitting important nuances that can

only be discovered through close reading. Therefore, when possible, analyses must be interpreted in the context of theories, historical data, events, and analyses of other kinds.

Future of Macroscope

Neither Tomas Engelthaler nor I is a professional computer scientist, therefore the Macroscope website is currently functioning but still has quite some room to improve before becoming a mature product. We will improve the Macroscope website in the coming six months. We have hired a part-time programmer to optimise the backend code. Analysis of all words will be pre-processed and stored so that users can immediately see the results after sending a query (currently a user must wait around 15 seconds for analysis of one word). A public API will be available to allow data retrieval without the need to use the website interface. Lastly, the website will be able to process large numbers queries at the same time. We have been talking with Oxford English Dictionary (OED) over the past few month in the hope to reaching cooperation of some kind in the future. With more influential parties involved, we may have greater resources to scale up functions and user-friendliness of the Macroscope.

Future direction of language evolution

The Macroscope provides a solid base to lead into further explorations on how language evolves. For example, how do word meaning change over time in relation to other word properties, such as concreteness, contextual diversity, frequency (Zipf, 1949), and age of acquisition (Kuperman, Stadthagen-Gonzalez, and Brysbasert 2012)? What are the additional structural properties of language that are associated with the birth and death processes of words (Pagel et al., 2007; Vejdemo & Hörberg, 2016)? Has language become more abstract through metaphor (Lakoff & Johnson, 1980) or become more concrete for easier acquisition (Hills & Adelman, 2015)?

A potential starting point could be a network model representing how words are structured according to their co-occurrence relationships. We have described in the chapter 1 that abstract words cannot attain their meaning through perceptual grounding, instead their meaning needs to be derived from their relationships with more concrete words whose meaning can be grounded in experience. This process may be reflected in a global network of words, which represents a language learning environment that the language users are exposed to. We can explore word properties (frequency, valence, concreteness, contextual diversity, etc) in relation to their roles in the global network. A few hypotheses could be:

1. The global network of words may demonstrate a radial structure: concrete, high-

frequent, short words cluster in the centre. These words are learnt first and through connections with them, meaning spreads towards the more peripheral space of the language structure. We may use the spreading activation model (Collins & Loftus, 1975) to model that process and test it against empirical data such as age of acquisition (Kuperman et al, 2012). The global network can be derived from either natural language data such as the Google Ngram Books, or mental association task (e.g. De Deyne et al, 2018; Dubossarsky, De Deyne, & Hills, 2017).

2. A global network of words may allow us to quantify the importance of each word to the stability of the global structure. Words with less connections with other words, positioned at peripheral area of the network, and used less frequently may hardly alter language structure if they were taken out of the system. Words that were important to the stability of global network may have experienced less semantic or frequency change.
3. We may also use the network to estimate the perceived concreteness of words. The idea is that if concrete words are the base for more abstract words to acquire their meaning, the concreteness of words should share certain network properties such as centrality. The historical trend of words becoming increasingly concrete/abstract may be described in terms of changing network structure. Abstract words may be perceived as more concrete over time through their connections with larger number of concrete words. For example, metaphorical expression that TIME IS MONEY makes *time* more concrete because this metaphor frames our experience with time in the context of money: we can *spend, save, spare, borrow, give* time just like what we do with money.

Lastly, I think a dataset that offers diachronic network properties of each word in the global language structure is able to inform language studies in learning, processing and evolution.

Envoi

One difficulty in the psychology lies in the familiarity of the phenomena with which it deals. Phenomena can be so familiar that we do not see them at all. This is especially true with language: our ability to learn language is endowed since the time of birth; our navigation through daily life depends on successful language processing; and our mental and physical world is contracted through language. This makes studying language challenging yet charming. With unprecedented large size of corpora and advances in natural language processing, I am

fortunate enough to study this ancient topic with new perspectives and methods. I hope a more quantitative approach to text analysis provides greater objectivity when dealing with a phenomenon so intertwined into our everyday life.

References

Alexander, M. G., Brewer, M. B., & Herrmann, R. K. (1999). Attitudes and affect: A functional analysis of out-group stereotypes. *Journal of Personality and Social Psychology, 77*(1), 78–93.

Alhothali, A., & Hoey, J. (2015). Good news or bad news: Using affect control theory to analyze readers' reaction towards news articles. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 1548–1558). Denver, CO: The Association for Computational Linguistics.

Allport, G. W. (1954). *The nature of prejudice*. New York: Addison.

Altarriba, J., & Bauer, L. M. (2004). The distinctiveness of emotion concepts: A comparison between emotion, abstract, and concrete words. *The American Journal of Psychology*, *117*(3), 389-410.

Anderson, J. R., & Bower, G. H. (1972). Recognition and retrieval processes in free recall. *Psychological Review*, *79*(2), 97-123.

Anthony, P. (1974) *The macroscope*. Sphere.

Arun, R., Suresh, V., Madhavan, C. V., & Murthy, M. N. (2010, June). On finding the natural number of topics with latent dirichlet allocation: Some observations. *The Pacific-Asia conference on knowledge discovery and data mining* (pp. 391-402). Berlin, Heidelberg: Springer.

Badley, M. M., & Lang, P. J. (1999). Affective norms for English words (ANEW): Instruction manual and affective ratings (Tech. Rep. No. C-1). Gainesville, FL: University of Florida, The Center for Research in Psychophysiology.

Bahrick, H. P. (1970). Two-phase model for prompted recall. *Psychological Review*, *77*(3), 215-222.

Barrett, L. F. (2006). Solving the emotion paradox: Categorization and the experience of emotion. *Personality and Social Psychology Review*, *10*(1), 20-46.

Barrett, L. F. (2017). *How emotions are made: The secret life of the brain*. Houghton Mifflin Harcour.

Barsalou, L. W. (2003). Situated simulation in the human conceptual system. *Language and Cognitive Processes, 18*(5-6), 513-62.

Barsalou, L. W. (2010). Grounded cognition: Past, present, and future. *Topics in Cognitive Science, 2*(4), 716–724.

Barsalou, L. W., Santos, A., Simmons, W. K., & Wilson, C. D. (2008). Language and simulation in conceptual processing. In M. De Vega, A. M. Glenberg, & A. C. A. Graesser (Eds.), *Symbols, embodiment, and meaning* (pp. 245–283). Oxford, England: Oxford University Press.

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology*, *5*(4), 323-370.

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). Beck depression inventory-II. *San Antonio*, *78*(2), 490-498.

Beck, U. (1992). *Risk society: Towards a new modernity.* London: Sage.

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: Analyzing text with the natural language toolkit*. Newton, MA: O'Reilly Media.

Blais, A.R., & Weber, E.U. (2006). A domain-specific ris-taking (DOSPERT) scale for adult populations. *Judgment and Decision Making, 1*(1)*,* 33–47.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*(Jan), 993-1022.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*(Jan), 993-1022.

Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, *2008*(10), P10008.

Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Advances in Neural Information Processing Systems* (pp. 4349-4357).

Borjas, G. J. (1990). *Friends or strangers: the impact of immigrants on the US economy*. New York: Basic Books.

Bourke, J. (2005). *Fear: A cultural history.* London: Virago.

Bower, G. H. (1981). Mood and memory. *American Psychologist*, *36*(2), 129-148.

Bradley, M. M., & Lang, P. J. (1999). Affective norms for English words (ANEW): Instruction manual and affective ratings (Tech. Rep. No. C-1). Gainesville, FL: University of Florida, The Center for Research in Psychophysiology.

Brewer, M. B. (1979). In-group bias in the minimal intergroup situation: A cognitive-motivational analysis. *Psychological Bulletin, 86*(2), 307-324.

Broadie, S., & Rowe, C. (2002). *Aristotle, Nicomachean ethics: translation, introduction and commentary.* Oxford: Oxford University Press.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 46(3), 904-911.

Bullinaria, J. A., & Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, *39*(3), 510-526.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science, 356*(6334), 183-186.

Carens, J. (2013). *The ethics of immigration*. New York: Oxford University Press.

Centers for Disease Control and Prevention (2010) *HIV/AIDS Surveillance Report*. Atlanta, GA: CDC.

Clark, G. (2008). *A farewell to alms: a brief economic history of the world*. Princeton University Press.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, *82*(6), 407-428.

Connell, L., & Lynott, D. (2011). Modality switching costs emerge in concept creation as well as retrieval. Cognitive Science, 35(4), 763-778.

Connell, L., & Lynott, D. (2016). Do we know what we're simulating? Information loss on transferring unconscious perceptual simulation to conscious imagery. Journal of Experimental Psychology: Learning, Memory, and Cognition, 42(8), 1218.

Cuddy, A. J. C., Fiske, S. T., Demoulin, S., & Leyens, J.-P. (2000). Stereotype content of social groups, as perceived by Belgian respondents. Unpublished data, Princeton University.

Cuddy, A. J., Fiske, S. T., & Glick, P. (2007). The BIAS map: Behaviors from intergroup affect and stereotypes. Journal of Personality and Social Psychology, 92(4), 631-648.

Damasio, A. R. (1989). Time-locked multiregional retroactivation: A systems-level proposal for the neural substrates of recall and recognition. *Cognition*, *33*(1-2), 25-62.

Darwin, C. (1872). *The expression of the emotions in man and animals*. London: Murray.

De Deyne, S., Navarro, D. J., Perfors, A., Brysbaert, M., & Storms, G. (2018). The "Small World of Words" English word association norms for over 12,000 cue words. *Behavior Research Methods. DOI 10.3758/s13428-018-1115-7*.

Deacon, T. W. (1997). *The symbolic species: The co-evolution of language and the brain*. New York: WW Norton & Company.

Delbecq-Derouesne, J., Beauvois, M. F., & Shallice, T. (1990). Preserved recall versus impaired recognition: A case study. *Brain*, *113*(4), 1045-1074.

Diener, E. D., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The satisfaction with life scale. *Journal of Personality Assessment*, *49*(1), 71-75.

Diener, E., Wirtz, D., Biswas-Diener, R., Tov, W., Kim-Prieto, C., Choi, D., et al. (2009). New measures of well-being. The Netherlands: Springer. The collected works of Ed Diener.

Diener, E., Wirtz, D., Tov, W., Kim-Prieto, C., Choi, D. W., Oishi, S., & Biswas-Diener, R. (2010). New well-being measures: Short scales to assess flourishing and positive and negative feelings. *Social Indicators Research*, *97*(2), 143-156.

Dixon, T. L. (2008). Crime news and racialized beliefs: Understanding the relationship between local news viewing and perceptions of African Americans and crime. *Journal of Communication, 58*(1), 106-125.

Dodds, P. S., Clark, E. M., Desu, S., Frank, M. R., Reagan, A. J., Williams, J. R., ... Megerdoomian, K. (2015). Human language reveals a universal positivity bias. *Proceedings of the National Academy of Sciences*, *112*, 2389–2394.

Douglas, M. (1992). *Risk and blame: Essays in cultural theory*. Abingdon, UK: Routledge.

Douglas, M., & Wildavsky, A. (1983). *Risk and culture: An essay on the selection of technological and environmental dangers.* Oakland, CA: University of California Press.

Dubossarsky, H., De Deyne, S., & Hills, T. T. (2017). Quantifying the structure of free association networks across the life span. *Developmental psychology*, *53*(8), 1560.

Dunbar, R. I. M. (1996). *Grooming, gossip, and the evolution of language*. London : Faber and Faber.

Eckes, T. (2002). Paternalistic and envious gender stereotypes: Testing predictions from the stereotype content model. *Sex Roles, 47*(3-4), 99-114.

Eichstaedt, J. C., Schwartz, H. A., Kern, M. L., Park, G., Labarthe, D. R., Merchant, R. M., ... & Weeg, C. (2015). Psychological language on Twitter predicts county-level heart disease mortality. *Psychological science*, *26*(2), 159-169.

Eichstaedt, J. C., Smith, R. J., Merchant, R. M., Ungar, L. H., Crutchley, P., Preoţiuc-Pietro, D., ... & Schwartz, H. A. (2018). Facebook language predicts depression in medical records. *Proceedings of the National Academy of Sciences*, *115*(44), 11203-11208.

Eisenstein, E. L. (1980). *The printing press as an agent of change* (Vol. 1). Cambridge University Press.

Ekman, P. (1992). An argument for basic emotions. *Cognition & Emotion*, *6*(3-4), 169-200.

Engelthaler, T., & Hills, T. T. (2018). Humor norms for 4,997 English words. *Behavior Research Methods,* 50(3), 1116-1124.

Estes, W., & DaPolito, F. (1967). Independent variation of information storage and retrieval processes in paired-associate learning. *Journal of Experimental Psychology, 75*(1), 18-26.

Fennelly, K., & Federico, C. (2008). Rural residence as a determinant of attitudes toward US immigration policy. *International Migration*, *46*(1), 151-190.

Ferreira, F., Bailey, K. G., & Ferraro, V. (2002). Good-enough representations in language comprehension. *Current Directions in Psychological Science, 11*(1), 11–15.

Ferrer-i-Cancho, R., & Vitevitch, M. S. (2017). The origins of Zipf's meaning-frequency law. arXiv preprint arXiv:1801.00168.

Fillmore, C. J. (1975). An alternative to checklist theories of meaning. In *Annual Meeting of the Berkeley Linguistics Society* (Vol. 1, pp. 123-131).

Fillmore, C. J. (1976). Frame semantics and the nature of language. *Annals of the New York Academy of Sciences*, *280*(1), 20-32.

Firth, J. R. (1957). *Papers in Linguistics 1934-1951.* London: Oxford University Press.

Fiske, S. T., Cuddy, A. J., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences, 11*(2), 77-83.

Fiske, S. T., Cuddy, A. J., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology, 82*(6), 878.

Fontaine, J. R., Scherer, K. R., Roesch, E. B., & Ellsworth, P. C. (2007). The world of emotions is not two-dimensional. *Psychological Science*, *18*(12), 1050-1057.

Freud, S. (1901). *Psychopathology of everyday life*. New York: Basic Books.

Fydrich, T., Dowdall, D., & Chambless, D. L. (1992). Reliability and validity of the Beck Anxiety Inventory. *Journal of Anxiety Disorders*, *6*(1), 55-61.

Gaissmaier, W., & Gigerenzer, G. (2012). 9/11, Act II: A fine-grained analysis of regional variations in traffic fatalities in the aftermath of the terrorist attacks. *Psychological Science, 23*(12), 1449–1454.

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, *115*(16), E3635-E3644.

Giddens, A. (1990). *The consequences of modernity*. Cambridge, England: Polity Press.

Glenberg, A. M. (1997). What memory is for. *Behavioral and Brain Sciences*, *20*(1), 1-19.

Glenberg, A. M., & Kaschak, M. P. (2002). Grounding language in action. *Psychonomic Bulletin & Review*, *9*, 558–565.

Greenfield, P. M. (2013). The changing psychology of culture from 1800 through 2000. *Psychological science*, *24*(9), 1722-1731.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National academy of Sciences*, *101*(suppl 1), 5228-5235.

Hall, L., Strandberg, T., Pärnamets, P., Lind, A., Tärning, B., & Johansson, P. (2013). How the polls can be both spot on and dead wrong: Using choice blindness to shift political attitudes and voter intentions. *PloS one, 8*(4), e60554.

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. arXiv preprint arXiv:1605.09096.

Hanley, J. R., Davies, A. D., Downes, J. J., & Mayes, A. R. (1994). Impaired recall of verbal material following rupture and repair of an anterior communicating artery aneurysm. *Cognitive Neuropsychology*, *11*(5), 543-578.

Hansen, C. H., & Shantz, C. A. (1995). Emotion-specific priming: Congruence effects on affect and recognition across negative emotions. *Personality and Social Psychology Bulletin*, *21*(6), 548-557.

Harari, Y. N. (2014). *Sapiens: A brief history of humankind*. New Tork: Random House.

Harari, Y. N. (2016). *Homo deus: A brief history of tomorrow*. New York: Random House.

Hartley, L.P. (1953). The Go-Between. Hamish Hamilton

Heider, E. R. (1972). Universals in color naming and memory. *Journal of Experimental Psychology*, *93*(1), 10.

Heider, E. R., & Olivier, D. C. (1972). The structure of the color space in naming and memory for two languages. *Cognitive Psychology*, *3*(2), 337-354.

Hills, T. T., & Adelman, J. S. (2015). Recent evolution of learnability in American English from 1800 to 2000. *Cognition*, *143*, 87-92.

Hills, T. T., Adelman, J. S., & Noguchi, T. (2016). Attention economies, information crowding, and language change. In Jones, M. N. (Ed.), *Big Data in Cognitive Science*. Psychology Press.

Hills, T. T., Jones, M. N., & Todd, P. M. (2012). Optimal foraging in semantic memory. *Psychological review*, *119*(2), 431-440.

Hills, T. T., Mata, R., Wilke, A., & Samanez-Larkin, G. R. (2013). Mechanisms of age-related decline in memory search across the adult life span. *Developmental psychology*, *49*(12), 2396.

Hills, T., Proto, E., & Sgroi, D. (2015) Historical analysis of national subjective wellbeing using millions of digitized books. *IZA Discussion Paper No. 9195.*

Hockett, C. F. (1960). The origin of speech. *Scientific American*, 203, 5–12.

Holzmann, R., & Jørgensen, S. (2001). Social risk management: A new conceptual framework for social protection, and beyond. *International Tax and Public Finance*, *8*(4), 529-556.

Hornik, K., & Grün, B. (2011). Topic models: An R package for fitting topic models. *Journal of Statistical Software*, *40*(13), 1-30.

Huntington, S. P. (1993). The clash of civilizations?. *Foreign Affairs*, *72*(3), 22-49.

Ito, T. A., Larsen, J. T., Smith, N. K., & Cacioppo, J. T. (1998). Negative information weighs more heavily on the brain: the negativity bias in evaluative categorizations. *Journal of Personality and Social Psychology*, *75*(4), 887-990.

Jagiello, R. D., & Hills, T. T. (2018). Bad news has wings: Dread risk mediates social amplification in risk communication. *Risk Analysis*, *38*(10), 2193-2207

Jeffers, R. J., & Lehiste, I. (1979). *Principles and methods for historical linguistics*. MIT press.

Jones, M. N., & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite. *Psychological review, 104, 1-37.*

Kasperson, R. E., Renn, O., Slovic, P., Brown, H. S., Emel, J., Goble, R., Kasperson, J. X., & Ratick, S. (1988). The social amplification of risk: A conceptual framework. *Risk Analysis*, *8*(2), 177-187.

Katz, J.J., & Fodor, J.A. (1963). The structure of a semantic theory. *Language, 39*(2), 170–210.

Kintsch, W. (1970). *Learning, memory, and conceptual processes*. New York: Wiley.

Koriat, A., & Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review*, *103*(3), 490.

Kramsch, C., & Widdowson, H. G. (1998). *Language and culture*. UK: Oxford University Press.

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, *44*(4), 978-990.

Ladd, D. R., Roberts, S. G., & Dediu, D. (2015) Correlational studies in typological and historical linguistics. Annual Review of Linguistics, 1 4.1–4.21.

Lakoff, G., & Johnson, M. (2008). *Metaphors we live by*. Chicago: University of Chicago Press.

Lampos, V., & Cristianini, N. (2012). Nowcasting events from the social web with statistical learning. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *3*(4), 72.

Landauer, I.K., & Dumais, S.T. (1997). A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*(2), 211-40.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse processes*, *25*(2-3), 259-284.

Le, X., Lancashire, I., Hirst, G., & Jokel, R. (2011). Longitudinal detection of dementia through lexical and syntactic changes in writing: a case study of three British novelists. *Literary and Linguistic Computing*, *26*(4), 435-461.

Levy, J.P., Bullinaria, J.A. & McCormick, S. (2017). Semantic Vector Evaluation and Human Performance on a New Vocabulary MCQ Test. In: Proceedings of the Thirty-ninth Annual Conference of the Cognitive Science Society, 2549-2554. Austin, TX: Cognitive Science Society.

Levy, O., Goldberg, Y., & Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, *3*, 211-225.

Li, Y., Engelthaler, T., Siew, C. S., & Hills, T. T. (2019). The Macroscope: A tool for examining the historical structure of language. *Behavior Research Methods,* 1-14.

Lin, Y., Michel, J., Aiden, E., Orwant, J., Brockman, W., & Petrov, S. (2012). Syntactic Annotations for the Google Books Ngram Corpus. *Proceedings of the ACL 2012 System Demonstrations* (pp 169–174). Association for Computational Linguistics.

Lindquist, K. A., & Barrett, L. F. (2008). Emotional complexity. In M. Lewis, J. Haviland-Jones & L. Feldman Barrett (Eds.), *Handbook of emotions* (3rd ed.), 513–530. New York, NY: Guilford Press.

Loewenstein, G. F., Weber, E. U., Hsee, C. K., & Welch, N. (2001). Risk as feelings. *Psychological Bulletin*, *127*(2), 267-286.

Louwerse, M. M. (2018). Knowing the Meaning of a Word by the Linguistic and Perceptual Company It Keeps. *Topics in Cognitive Science, 10*(3): 573-589.

Lovibond, P. F., & Lovibond, S. H. (1995). The structure of negative emotional states: Comparison of the Depression Anxiety Stress Scales (DASS) with the Beck Depression and Anxiety Inventories. *Behaviour Research and Therapy*, *33*(3), 335-343.

Lucas, R. E., Diener, E., & Larsen, R. J. (2003). Measuring positive emotions. In S. J. Lopez & C. R. Snyder (Eds.), *Positive psychological assessment: A handbook of models and measures* (pp. 201–218). Washington, DC: American Psychological Association.

Malt, B. C., & Smith, E. E. (1984). Correlated properties in natural categories. *Journal of Memory and Language*, *23*(2), 250.

Markowitz, H. (1952). Portfolio selection. *The Journal of Finance*, *7*(1), 77–91.

McDowell, I. A. N., & Praught, E. D. (1982). On the measurement of happiness: an examination of the Bradburn Scale in the Canada Health Survey. *American Journal of Epidemiology*, *116*(6), 949-958.

Mervis, C. B., & Rosch, E. (1981). Categorization of natural objects. *Annual Review of Psychology*, *32*(1), 89-115.

Michel, J. B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., ... & Pinker, S. (2011). Quantitative analysis of culture using millions of digitized books. *science*, *331*(6014), 176-182.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111-3119).

Monaghan, P. (2014). Age of acquisition predicts rate of lexical evolution. Cognition, 133(3), 530-534.

Morgan, C. D., & Murray, H. A. (1935). A method for investigating fantasies: The Thematic Apperception Test. *Archives of Neurology & Psychiatry*, *34*(2), 289-306.

Moussaïd, M., Brighton, H., & Gaissmaier, W. (2015). The amplification of risk in experimental diffusion chains. *Proceedings of the National Academy of Sciences*, *112*(18), 5631-5636.

Oeppen, J., & Vaupel, J. W. (2002). Broken limits to life expectancy. *Science*, *296*(5570), 1029–1031.

Oishi, S., Schimmack, U., & Colcombe, S. J. (2003). The contextual and systematic nature of life satisfaction judgments. *Journal of Experimental Social Psychology*, *39*(3), 232-247.

Osgood, C., Suci, G., & Tannenbaum, P. (1957). *The measurement of meaning*. Urbana, IL: University of Illinois.

Pagel, M., Atkinson, Q. D., & Meade, A. (2007). Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. *Nature, 449*(7163), 717.

Paivio, A. (1971). *Imagery and verbal processes.* New York: : Holt, Rinehart, & Winston,

Paivio, A. (1986). *Mental representations: A dual coding approach.* New York: Oxford University Press.

Paivio, A., Yuille, J. C., & Madigan, S. A. (1968). Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of Experimental Psychology, 76,* 1–25.

Peabody, D. (1985). *National characteristics*. New York: Cambridge University Press.

Pechenick, E. A., Danforth, C. M., & Dodds, P. S. (2015). Characterizing the Google Books corpus: Strong limits to inferences of socio-cultural and linguistic evolution. *PloS One*, *10*(10), e0137041.

Pecher, D., & Zwaan, R. A. (Eds.) (2005). *Grounding cognition: The role of perception and action in memory, language, and thinking*. New York: Cambridge University Press.

Peirce, C. S. (1931). Collected papers of Charles Sanders Peirce. Cambridge MA: Harvard University Press.

Pennebaker, J. W., & Stone, L. D. (2003). Words of wisdom: language use over the life span. *Journal of personality and social psychology*, *85*(2), 291.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532-1543).

Petersen, A. M., Tenenbaum, J. N., Havlin, S., Stanley, H. E., & Perc, M. (2012). Languages cool as they expand: Allometric scaling and the decreasing need for new words. *Scientific reports*, *2*, 943.

Pettigrew, T. F. (1979). The ultimate attribution error: Extending Allport's cognitive analysis of prejudice. *Personality and Social Psychology Bulletin, 5*(4), 461-476.

Pettigrew, T. F., & Tropp, L. R. (2006). A meta-analytic test of intergroup contact theory. *Journal of Personality and Social Psychology*, *90*(5), 751-783.

Piketty, T. (2014). Capital in the 21st century. Harvard University Press.

Pinker, S. (2011). *The better angels of nature: The decline of violence in history and its causes*. London: Penguin.

Pleskac, T. J., & Hertwig, R. (2014). Ecologically rational choice and the structure of the environment. *Journal of Experimental Psychology: General*, *143*(5), 2000-2019.

Poppe, E. (2001). Effects of changes in GNP and perceived group characteristics on national and ethnic stereotypes in Central and Eastern Europe. *Journal of Applied Social Psychology, 31*(8), 1689–1708.

Portes, A., & Sensenbrenner, J. (1993). Embeddedness and immigration: Notes on the social determinants of economic action. *The American Journal of Sociology, 98*(6), 1320–1350.

Portes, A., & Zhou, M. (1993). The new second generation: Segmented assimilation and its variants. *The Annals of the American Academy of Political and Social Science, 530*(1), 74–96.

Pratt, J.W. (1964). Risk aversion in the small and in the large. *Econometrica*, *32*(1/2), 122–136.

Pulvermuller, F., & Pulvermüller, F. (2002). The neuroscience of language: On brain circuits of words and serial order. Cambridge University Press.

Raaijmakers, J. G., & Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review*, *88*(2), 93-134.

Ramscar, M., Hendrix, P., Shaoul, C., Milin, P., & Baayen, H. (2014). The myth of cognitive decline: Non-linear dynamics of lifelong learning. *Topics in cognitive science*, *6*(1), 5-42.

Rayner, S., & Cantor, R. (1987). How Fair Is Safe Enough? The Cultural Approach to Societal Technology Choice 1. *Risk Analysis*, *7*(1), 3-9.

Recchia, G., & Louwerse, M. M. (2015). Reproducing affective norms with lexical co-occurrence statistics: Predicting valence, arousal, and dominance. *The Quarterly Journal of Experimental Psychology, 68*(8), 1584-1598.

Robinson, M. D., & Johnson, J. T. (1996). Recall memory, recognition memory, and the eyewitness confidence–accuracy correlation. *Journal of Applied Psychology*, *81*(5), 587.

Rorschach, H. (1921). *Psychodiagnostik*. Leipzig, Germany: Ernst Bircher Verlag.

Rosch, E. (1973). On the internal structure of perceptual and semantic categories. In T. E. Moore (Ed.), *Cognitive development and the acquisition of language*. New York: Academic Press.

Rosch, E. (1973). Natural categories. *Cognitive Psychology*, *4*(3), 328-350.

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology,104* (3), 192–233.

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology, 7*(4), 573–605.

Rozin, P., & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review*, *5*(4), 296-320.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*(6), 1161-1178.

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological Review*, *110*(1), 145–172.

Russell, J. A., & Barrett, L. F. (1999). Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant. *Journal of Personality and Social Psychology*, *76*(5), 805.

Ryff, C. D., & Keyes, C. L. M. (1995). The structure of psychological well-being revisited. *Journal of personality and social psychology*, *69*(4), 719-727.

Sagi, E., Kaufmann, S., & Clark, B. (2011). Tracing semantic change with latent semantic analysis. *Current Methods in Historical Semantics, 73*,161-183.

Salovey, P., & Mayer, J. D. (1990). Emotional intelligence. *Imagination, Cognition and Personality*, *9*(3), 185-211.

Sandhaus E (2008) The New York Times Annotated Corpus. (Linguistic Data Consortium, Philadelphia).

Schönemann, P. H. (1966). A generalized solution of the orthogonal procrustes problem. *Psychometrika*, *31*(1), 1-10.

Schwanenflugel, P. (1991). Why are abstract concepts hard to understand? In P. J. Schwanenflugel (Ed.), The psychology of word meanings (pp. 223–250). Hillsdale, NJ: Erlbaum.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., ... & Ungar, L. H. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, *8*(9), e73791.

Schwarz, N., Bless, H., & Bohner, G. (1991). Mood and persuasion: Affective states influence the processing of persuasive communications. *Advances in Experimental Social Psychology*, *24*, 161-199.

Scollon, C. N., Diener, E., Oishi, S., & Biswas-Diener, R. (2004). Emotions across cultures and methods. *Journal of Cross-cultural Psychology*, *35*(3), 304-326.

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, *3*(3), 417-424.

Semin, G. R., & Smith, E. R. (Eds.) (2008). *Embodied grounding: Social, cognitive, affective, and neuroscientific approaches*. Cambridge, MA: Cambridge University Press.

Shapiro, L. (2014). *The Routledge handbook of embodied cognition*. London: Routledge.

Sherif, M., Harvey, O. J., White, B. J., Hood, W. R., & Sherif, C. W. (1961). *Intergroup conflict and cooperation: The Robbers Cave experiment.* Norman, OK: University Book Exchange.

Short, J. F. (1984). The social fabric at risk: toward the social transformation of risk analysis. *American sociological review, 49*(6), 711-725.

Sievert, C., & Shirley, K.E. (2014). LDAvis: A method for visualizing and interpreting topics. *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, pp. 63–70. Association for Computational Linguistics, Baltimore, Maryland, USA.

Simon, H. (1991). Designing organisations for an information-rich word. In M, Greenberger (Eds.), *Computers, communications, and the public interest* (pp. 40-41). Baltimore, MA: The Johns Hopkins Press.

Simon, H. A. (1991). The architecture of complexity. In *Facets of systems science* (pp. 457-476). Springer, Boston, MA.

Skeldon, R. (2014). *Migration and development: A global perspective*. New York: Routledge.

Slovic, P. (1987). Perception of risk. *Science*, *236*(4799), 280–285.

Slovic, P., Fischhoff, B., & Lichtenstein, S. (1985). Characterizing perceived risk. In W. Kates, C. Hohenemser, & J. X. Kasperson (Eds.), *Perilous progress: Managing the hazards of technology* (pp. 91-125). Boulder, CO: Westview.

Snefjella, B., & Kuperman, V. (2015). Concreteness and psychological distance in natural language use. *Psychological Science, 26*(9), 1449-1460.

Sunstein, C. R. (2005). *Laws of fear: Beyond the precautionary principle*. Cambridge, UK: Cambridge University Press.

Tabor, D., & Stockley, L. (2018). *Personal well-being in the UK: January to December 2017.* UK: Office for National Statistics.

Taddy M (2012) On estimation and selection for topic models. *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics* (pp 1184–1193).

Tajfel, H. (1982). Social psychology of intergroup relations. *Annual Review of Psychology, 33*(1), 1-39.

Tajfel, H., Billig, M. G., Bundy, R. P., & Flament, C. (1971). Social categorization and intergroup behaviour. *European Journal of Social Psychology*, *1*(2), 149-178.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, *29*(1), 24-54.

Thomas, D. L., & Diener, E. (1990). Memory accuracy in the recall of emotions. *Journal of Personality and Social Psychology*, *59*(2), 291.

Thompson, E. R. (2007). Development and validation of an internationally reliable short-form of the positive and negative affect schedule (PANAS). *Journal of Cross-cultural Psychology, 38*(2), 227-242.

Trope, Y., & Liberman, N. (2003). Temporal construal. *Psychological Review, 110*(3), 403-421.

Trope, Y., & Liberman, N. (2010). Construal-level theory of psychological distance. *Psychological Review, 117*(2), 440-463.

Tugade, M. M., Fredrickson, B. L., & Feldman Barrett, L. (2004). Psychological resilience and positive emotional granularity: Examining the benefits of positive emotions on coping and health. *Journal of Personality*, *72*(6), 1161-1190.

Tulving, E., & Pearlstone, Z. (1966). Availability versus accessibility of information in memory for words. *Journal of Verbal Learning and Verbal Behavior*, *5*(4), 381-391.

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37, 141-188.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, *5*(2), 207-232.

Twenge, J. M., Campbell, W. K., & Gentile, B. (2012). Increases in individualistic words and phrases in American books, 1960–2008. *PloS One*, *7*(7), e40181.

U.S. Department of Homeland Security. (2010). Yearbook of Immigration Statistics 2017. Retrieved from: https://www.dhs.gov/immigration-statistics.

Uz, I. (2014). Individualism and first person pronoun use in written texts across languages. *Journal of Cross-Cultural Psychology*, *45*(10), 1671-1678.

Van Rensbergen, B., Kuppens, P., Storms, G., & De Deyne, S. (2015). *Computationally coding responses of a free-format self-description personality test using word association data.* (Doctoral dissertation).

Vejdemo, S., & Hörberg, T. (2016). Semantic factors predict the rate of lexical replacement of content words. *PloS one*, *11*(1), e0147924.

Wang, Z., Busemeyer, J. R., Atmanspacher, H., & Pothos, E. M. (2013). The potential of using quantum theory to build models of cognition. *Topics in Cognitive Science*, *5*(4), 672-688.

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, 45(4), 1191-1207.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of Personality and Social Psychology*, *54*(6), 1063.

Westbury, C., Keith, J., Briesemeister, B. B., Hofmann, M. J., & Jacobs, A. M. (2015). Avoid violence, rioting, and outrage; approach celebration, delight, and strength: Using large text corpora to compute valence, arousal, and the basic emotions. *The Quarterly Journal of Experimental Psychology*, *68*(8), 1599-1622.

Wilson-Mendenhall, C. D., Simmons, W. K., Martin, A., & Barsalou, L. W. (2013). Contextual processing of abstract concepts reveals neural representations of nonlinguistic semantic content. *Journal of Cognitive Neuroscience*, *25*(6), 920-935.

Wittgenstein, L. (2009). *Philosophical investigations*. John Wiley & Sons.

Woods, J., & Arthur, C. D. (2017). *Debating immigration in the age of terrorism, polarization, and trump*. London: Lexington Books.

World Economic Forum (2017) *The Global Risks Report 2017, 12th Edition*. Geneva: World Economic Forum.

World Health Organization. (2009). *Global Health Risks: Mortality and Burden of Disease Attributable to Selected Major Risks*. Geneva: World Health Organization.Retrieved from: https://apps.who.int/iris/handle/10665/44203.

Wundt, W. (1905). *Fundamentals of physiological psychology*. *Leipzig: Engelmann*.

Xu, Y., & Kemp, C. (2015). A computational evaluation of two laws of semantic change. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (Eds.), Proceedings of the 37th Annual Meeting of the Cognitive Science Society (pp. 2703–2708). Austin, TX: Cognitive Science Society.

Yarkoni, T. (2010). Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers. *Journal of Research in Personality*, *44*(3), 363-373.

Zalizniak, A., Bulakh, M., Ganenkov, D., Gruntov, I., Maisak ,T., & Russo, M. (2012). The catalogue of semantic shifts as a database for lexical semantic typology. *Linguistics*, 50, 633–69.

Zevon, M. A., & Tellegen, A. (1982). The structure of mood change: An idiographic/nomothetic analysis. *Journal of Personality and Social Psychology*, *43*(1), 111.

Zipf, G. (1949). *Human behavior and the principle of least effort.* New York: Addison-Wesley.

Zwaan, R. A. (2014). Embodiment and language comprehension: Reframing the discussion. *Trends in Cognitive Sciences*, 18(5), 229–234.